THE NATIONAL UNIVERSITY
of SINGAPORE

School of Computing
Lower Kent Ridge Road, Singapore 119260

TRC9/06

Coronary Artery Disease Prediction with Bayesian Networks and Constraint Elicitation

Qiongyu CHEN, Guoliang LI, Bin HAN,
Chew Kiat, HENG and Tze-Yun LEONG

September 2006

Technical Report

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

JAFFAR, Joxan
Dean of School

Coronary Artery Disease Prediction with Bayesian Networks and Constraint Elicitation

Qiongyu Chen[1], Guoliang Li[1], Bin Han[1], Chew Kiat Heng[2], and Tze-Yun Leong[1]

[1] Medical Computing Laboratory, School of Computing
National University of Singapore
3 Science Drive 2, Singapore 117543
[2] Department of Paediatrics, National University of Singapore
5 Lower Kent Ridge Road, Singapore 119074

Abstract. Coronary artery disease (CAD) is one of the major causes of death in the world. Finding cost-effective methods to predict CAD is a major challenge for public health. In this paper, we propose a Bayesian network learning approach with constraint elicitation mechanism to predict the risk of CAD. The underlying causal assumption and interpretability make Bayesian networks a good tool for medical applications, in this case CAD risk prediction involving both genetic and environmental factors. The constraint elicitation process improves model accuracy by incorporating relevant domain knowledge. We performed experiments to compare our results with those from other machine learning methods, such as naive Bayes, support vector machines, K nearest neighbors, neural networks and decision trees. Our method is shown to be comparable to these methods in terms of prediction accuracy but at the same time offers an intuitive representation of the relationships among variables in the problem domain. Conforming to the domain knowledge, the results identified the important environmental factors for CAD prediction and the relevant groups of gene markers contributing to the risk of CAD. The results also indicated that some gene markers that are relevant to CAD risk in western populations, but may not be relevant in Chinese, Indian and Malay populations local to Singapore.

1 Introduction

Coronary artery disease (CAD) is the most common form of heart disease in America and Europe [1, 2]. It occurs when the arteries that supply blood to the heart muscles (coronary arteries) become hardened and narrowed. Every year, millions of deaths worldwide are attributed to CAD. Therefore, finding cost-effective methods to predict and control CAD is one of the greatest challenges in public health.

Research in this area usually involves using medical profile and family history information to predict the risk for CAD. For example, Lapuerta *et. al.*, [9] used seven different mean lipid values in neural networks to predict the occurrence of

a complication of coronary artery disease. Wilson *et. al.*, [12] predict the risk of CAD by identifying risk categories and statistical tests, e.g., linear regression and logistic regression.

Along with the rapid advancement of biomedical technologies, now we can combine genetic information, such as microarray-based genotyping, with clinical and environmental factors to predict the risk of CAD. This integrative approach can give us better understanding of the fundamentals of the disease. Tham *et. al.* [11] combined medical profile information, family history information and microarray-based genotyping information in a neural network to predict the risk of CAD, and achieved reasonably good results. However, the neural networks are a type of black-box method; the resulted networks and functions used in the method are not easily interpretable. This is often undesired in the medical domain.

In this paper, we propose to assess the risk of CAD with Bayesian networks. The graph structure in Bayesian networks is easy to interpret and may imply causal relationships. These properties make the method a good tool for medical applications, in this case CAD risk prediction involving both genetic and environmental factors. Our experiments show that the initial Bayesian Networks learned from data are not good enough for prediction purpose. However, after augmenting the Bayesian network learning with domain knowledge, the prediction accuracy increased significantly. We incorporated domain knowledge through our proposed interactive and iterative constraint elicitation process, which can minimize the expert's effort in domain knowledge elicitation. We performed experiments to compare our methods with other machine learning methods, such as naive Bayes, support vector machines, K nearest neighbors, neural networks and decision trees. Results show that our method is comparable to these methods in terms of prediction accuracy. In addition, the relationships among the variables are easily interpretable from the learned Bayesian network. Specifically, we observed a few interesting findings from the learned Bayesian network: the important factors to CAD, the possible groupings of the variables and the possible irrelevant gene markers to Asian populations.

2 Background

Bayesian networks are graphical tools for modeling uncertainty and inferring causal relationships among variables in an uncertain domain. A Bayesian network is represented as a directed acyclic graph (DAG) (which implies no feedback in the underlying domain). The nodes in the structure represent the variables in the domain. Arcs between nodes represent probabilistic dependencies between different variables, and often imply causal relationships. Together, the graphical representation models qualitative information, while the conditional probability table in each node models quantitative information in the domain.

Bayesian networks can be constructed entirely from domain knowledge. But the process is very time-consuming. To overcome this problem, there are many research efforts to build Bayesian networks from data [3–5, 7]. For example, to

learn the structure of a Bayesian Network, a family of score-based algorithms search through potential network spaces for the network with the best score using some pre-defined scoring function. The scoring function usually reflects how well a network topology fits the data set, e.g., Bayesian Information Criterion (BIC) [10]. Search methods may range from greedy search to exhaustive search, though the latter is not always feasible for Bayesian networks with a moderately-large number of variables.

Constraint-based algorithms infer network topology based on dependencies among variables. These dependencies are measured by some pre-defined statistical tests. The pairs of nodes with results above certain thresholds under certain conditions are considered to be dependent and an edge is added between them. Due to the construction method, a constraint-based algorithm often builds an undirected graph from the variables and resolves link direction at a later stage.

Some of these methods specifically target the classification problem. For example, Naive Bayes is a type of Bayesian networks targetted at the classification problem with strong assumptions and constraints - it assumes that the class label is the only root (a node without parents) in a Bayesian network, and all the other attributes are conditionally independent given the class label. Although Naive Bayes achieves reasonably good performance in many applications, its assumptions cannot be easily imposed in many real problems.

Friedman *et. al.*, [5] introduced a Tree Augmented Naive Bayes (TAN) method as a Bayesian network classifier. They assumed that the class label is the only root in the Bayesian networks and other variables have at most one other variables as parents besides the root. This is a significant improvement compared with Naive Bayes; however, the assumptions are still quite rigid and not generally applicable.

3 Learning Bayesian networks with constraint elicitation

Generally, the Bayesian networks learned from data are not good enough for classification purposes [5]. Relevant domain knowledge is very important for building Bayesian networks to support classification and prediction. However, building Bayesian networks entirely from domain knowledge is time-consuming. In order to maximize the usage of the data and minimize the experts' efforts to accurately model the relationships among the domain variables, we propose an iterative and interactive process to learn Bayesian networks from data and elicit constraints from domain experts based on the learned Bayesian networks.

In the process, we first learn an initial Bayesian network from data without any constraints. Then a domain expert will judge the significance of the following five types of constraints in the learned Bayesian network:

1. Whether the roots in the learned Bayesian network are reasonable. Roots are variables which influence other variables, but are not influenced by any other variables;

2. Whether the leaves in the learned Bayesian network are reasonable. Leaves are variables that are influenced by some variables, but do not influence any other variables;
3. Whether the direct links between some pairs of variables are deemed important by experts, conditioning on that the experts have concrete knowledge about the links;
4. Whether the causal order of two variables are correct. The causal order of two variables means which one variable comes before the other in terms of cause-effect relationship;
5. Whether the sub-groups of the original variables are reasonable. Some variables, especially those among genotype data, can be observed to form groups in the learned Bayesian networks.

Some of these constraints can be easily assessed by domain experts, while others may be more complicated. Usually it takes several rounds to assess the links between two variables. After each assessment, we learn a Bayesian network from data with the added constraints. After several iterations, when domain experts deem the learned Bayesian Network consistent with the existing biological and medical knowledge, we stop the learning process and measure the performance of the final Bayesian network using prediction accuracy. The whole interactive and iterative process to learn Bayesian networks with constraint elicitation is as follows:

- Step 1: Set the constraint set empty
- Step 2: Learn a Bayesian network from data with the current constraint set
- Step 3: Check whether there are new constraints from the learned Bayesian networks; The type of constraints is mentioned above
- Step 4: If there are new constraints, add the constraints to the constraint set, and go to Step 2
- Step 5: If there are no constraints, make prediction on the test data with the current Bayesian network

4 Experiments and discussion

4.1 The HEART data set

The data set used in this research is an expanded set from one which was originally collected from a Singaporean population for predicting CAD risk using neural networks [11]. It contains information from a total of 2,949 subjects, 41 variables from each profile: ten environmental risk factors, thirty genotyping information and one class label. The term "environmental risk factors" is used broadly here to include all risk factors that are not obtained from genotyping. Some are clearly environmental in nature, such as smoking; but it also includes traits that may be genetic in nature, such as ethnicity and family history of diseases. Eight of the environmental factors are discrete variables and two others are continuous variables. The distributions of the continuous variables, age and

body mass index (BMI), roughly follow the normal distribution. Both of them were discretized into 9 equal-width categories separately. The environmental risk factors are summarized in Table 1.

Node title	No. of states	Remarks
CAD	2	Healthy or diseased
SEX	2	Male or female
RACE	3	Chinese, Indian and Malay
DM	2	Diabetes, healthy or diseased
HY	2	Hypertension, healthy or diseased
SM	3	Smoker, non-smoker and ex-smoker
FCAD	2	Family history of CAD, yes or no
FDM	2	Family history of diabetes, yes or no
FHY	2	Family history of hypertension, yes or no
AGE	continuous	
BMI	continuous	body-mass index

Table 1. The summary of the data set

Each medical profile contains 30 candidate gene markers, or gene polymorphisms that may render the patient susceptible to developing coronary artery disease for Caucasians [8]. The influences of these genes on Asians (Chinese, Indians and Malays in this study) are unclear. The possible outcomes of these gene markers denote the genotypes of a particular gene, which is often a substitution of a nucleotide (e.g., T substituting C) at a certain position of the gene. In this data set, each gene marker has 3 possible genotypes. Although the specific functions of some of the 30 candidate genetic markers are known, many are still unclear. This makes it very impossible to study their etiological relationships with CAD.

The class label is CAD, with 0 denoting healthy subjects and 1 denoting patients suffering from CAD. We know that CAD is a disease spectrum, which can be categorized in terms of the number of the blocked vessels. In our work, we set the class label as 1 if the presence of at least 50% narrowing in at least one of the major coronary arteries in the subject as ascertained by angiography is observed. Such a categorization captures the distinction between healthy and disease samples. In the data, 1,462 out of the 2,949 subjects, or 49.6%, constituted the disease cases while the remaining, who were healthy at the time of recruitment, served as healthy controls.

4.2 The results at different stages of constraint elicitations

In our experiments we used constraint-based structure learning algorithm for Bayesian networks with a significance level of 0.05. All the results are based on 10 fold cross validation. Specifically, the entire data set was randomly split

into 10 approximately equal folds, and in turn each fold is used as test set and other folds are used as training set. The average of the ten prediction results are reported. The performances are measured by sensitivity, specificity and prediction accuracy. Sensitivity is defined as TP/(TP + FN), specificity as TN/(TN + FP), and prediction accuracy as (TP + TN)/(TP + FP + TN + FN), where TP, TN, FP and FN denote the numbers of the true positives, true negatives, false positives, and false negatives, respectively.

At the first stage, we learned a Bayesian network from the data without any constraints, which is shown in Figure 1. The classification accuracy is 79.42%, which is worse than other classifying algorithms. Also, some links among the variables are not consistent with domain knowledge. For example, there exist directed link between "Age" and "race", "sex". At the second stage, this preliminary Bayesian Network was presented to a domain expert for comments. Some variables were deemed to be roots or leaves in a Bayesian network. For example, the values of the variable "Age" change with time, and are believed not to be affected by other variables, for example medical profile, family history or genetic information. So it cannot be a child node of any other variables. The arcs going into variable "Age" may be due to data collection bias or wrongly induced arc direction. These arcs are subsequently either deleted or reversed. Some other patterns exhibited in the Bayesian Network were also identified for further study, for example, gene markers G8, G9 and G10, which are shown to be not related to CAD in our research population.

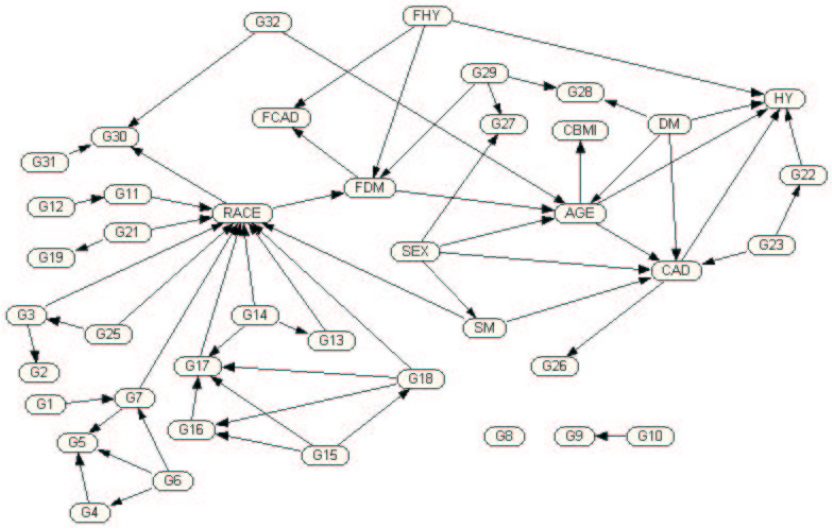

Fig. 1. The initial Bayesian Network learned from HEART data set

After several iterations, the constraints we took into consideration are: variables "Age" and "Sex" are roots in the learned Bayesian networks; no direct links among "Race", "Age", and "Sex"; and the direction of the arcs between CAD and some gene markers. When there are no new constraints available, the final Bayesian network is learned with constraint-based Bayesian network learning algorithm incorporated with the constraints from domain experts. The classification accuracy from the final Bayesian network has achieved a higher value of 86.49%. The classification accuracies for different stages are shown in Table 2.

In Table 2, the Bayesian network learned in the intermediate stage has one constraint incorporated – "age" is a root node. This increases the prediction accuracy. Adding more constraints, such as the variable "sex" being a root and that the arcs between gene markers and CAD should have directions from the former to the latter, the prediction rate further increases. Our result suggests that incorporating domain knowledge is essential for constructing Bayesian network structure. In addition, our interactive and iterative process ensures that domain experts spend minimum efforts to assess the constraints in the learned Bayesian networks. They may choose to selectively add in those constraints that are of interests, or about which there exists concrete knowledge. It is not necessary to assess the relationships among all possible pairs of variables in the domain, which is often the case for constructing Bayesian networks entirely from domain knowledge.

Algorithm	Sensitivity	Specificity	Accuracy
No constraints	0.8327	0.7683	0.7942
Intermediate stage	0.8851	0.8223	0.8535
Final graph	0.8993	0.8299	0.8649

Table 2. The results of Bayesian Networks prediction with different constraints

4.3 Different patterns identified in HEART data set

The final learned Bayesian network with constraints are shown in Figure 2. The relationships among the variables are consistent with domain knowledge. We performed sensitivity analysis by measuring the mutual information of each variable to node CAD from the learned Bayesian network. The results show that the top four risk factors for CAD are diabetes, hypertension, smoking status, and age, which conform to the existing medical knowledge. Also, from the domain knowledge, we know that no single gene makes a major contribution to the disease. This is consistent with our graph that no single gene marker is highly ranked in the sensitivity analysis. Moreover, we know that some environmental factors, such as race and related diseases, integrate some of the effects of gene polymorphisms, including the effects of their interactions with each. This may

contributes to the fact that the environmental factors have more discriminant power for CAD prediction.

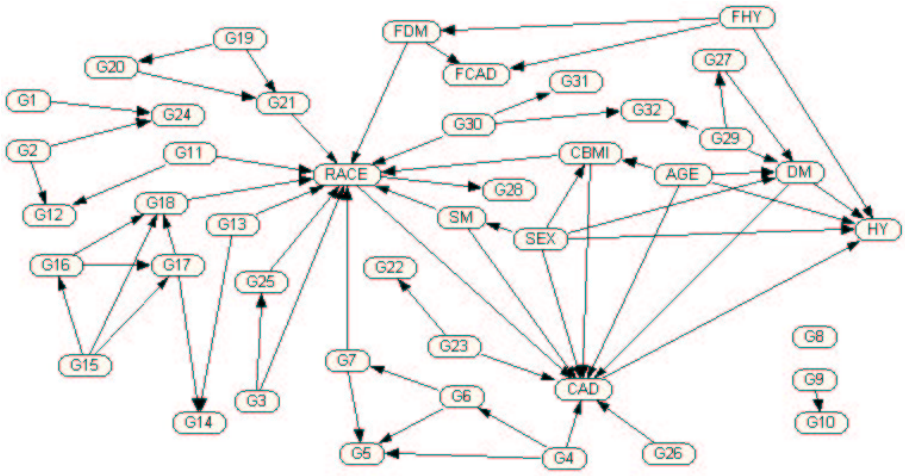

Fig. 2. The final Bayesian Network learned from HEART data set

From the final Bayesian network learned, we have also derived several interesting findings. Firstly, we observe that some genetic variables are related to each other and affect the classification accuracy as a group. For example, we observe that G13 to G18 are closed related to one another, and only one or two of the group are linked with other variables. After consulting with the domain expert, we found that such six gene markers are six different single nucleotide polymorphisms at different positions in one gene. The presence of one strongly influences the presence of another in the group, and their distributions are highly correlated with the race of the person. Secondly, we observe that some variables are not connected with CAD in the graph via any path, such as gene markers G8, G9 and G10. This disconnection suggests that these variables may be irrelevant to the CAD in the local Chinese, Indian and Malay populations, although such gene markers are relevant factors for the western populations. Such findings are deemed to warrant further verification with genetic association studies in the Singaporean population.

4.4 Comparison with other methods

Experiments show that our results are significatly better than those in a similar study using neural networks with committee learning [11]. We also compared the performance of our method with some other data mining methods: Decision trees, support vector machines, neural networks, k nearest neighbors, and naive

Bayes. The decision tree methods are a *de facto* classification method to evaluate other classification methods. It is built recursively based on the information gain of the features. Support vector machines (SVM) find the optimal decision plane by selecting the fewest instances as the support vectors with the largest margin in the feature space. It is probably the classification method with the best prediction results up to date, although it sometimes suffers in the presence of noisy data. Neural networks are a type of black-box method which can approximate any continuous function to arbitrary accuracy provided that the model has sufficiently large number of nodes and the parameters of the model are chosen properly. K-nearest neighbor classifiers are instance-based methods. They make predictions based on the distances between the test data and the training data. Naive Bayes is a probability-based classification method; it assumes that all the features are independent of each other given the class. It is a simple but practical classification method. Some of these algorithms, especially support vector machines, gave very good prediction results. However, most of these classifiers do not allow for meaningful interpretation of the relationship among variables in the problem domain, thus provide limited understanding into the problem domain.

We used the implementations of the above algorithms in the WEKA [13] software package. The results are shown in Table 3. We observe that the sensitivity, specificity and accuracy from our method are comparable to the other methods, which means that our learned model with high joint probability can achieve reasonable classification results. In addition, our method is able to identify relationship patterns among the variables without specifying a target variable in the learning process. This makes our method more suitable for hypotheses generation in knoweldge discovery.

Algorithm	Sensitivity	Specificity	Accuracy
Our method	0.8993	0.8299	0.8649
Tham *et. al.*	Unknown	Unknown	0.8075
Decision Tree (J48)	0.8722	0.8775	0.8749
K nearest neighbors	0.8762	0.7694	0.8233
Naive Bayes	0.8809	0.8905	0.8857
Support vector machine	0.8917	0.8878	0.8897
Neural Network (MLP)	0.8735	0.8768	0.8752

Table 3. Prediction results with different algorithms

5 Discussion and future work

Coronary artery disease (CAD) is one of the major causes of death in the world. Finding cost-effective methods to predict CAD is a major challenge for public

health. In this paper, we proposed a Bayesian network learning approach with constraint elicitation mechanism for CAD prediction and risk analysis, which takes advantage of both statistical learning and domain knowledge. Our method is particularly helpful when partial knowledge of the problem domain is known, but insufficient to construct a complete Bayesian network. We compared our method with other well-known classification algorithms. The comparison shows that our method is comparable to these methods in terms of classification accuracy. In addition, the resulting graphical representation is interpretable and understandable to domain experts. We also identified some relevant patterns in the variables. Some gene markers do not have significant correlation with CAD prediction based on the data set we used. For example, in our study gene markers G8, G9 and G10 have no direct arcs to the rest of the variables in the problem domain, indicating that they are not closely correlated with risk of CAD in our research population. Some variables are correlated in a group, like G13 to G18. These patterns warrant further investigation. Our future work includes futher developing the constraints evaluation process, as well as solving protential conflicts in the constraints elicitated. These conflicts may arise from differing domain experts' opinion, or change of opinion by the same domain expert after new medical evidences.

The combined effect of genetic and environmental factors on the risk of CAD has not been extensively studied in the medical domain. Although our results successfully identified genetic and environmental factors that have significant influence respectively on CAD, the interplay of these two groups of factors and their combined effects are still unclear. The Bayesian Network, being a flat structure with no structural uncertainty, may prove inadequate for this purpose. We're looking for other probabilistic graphical models with richer representation power, e.g., Probabilistic Relational Models [6], which may give us further insights into this problem.

6 Acknowledgments

This research was partially supported by Research Grant No. BM/00/007 from the Biomedical Research Council (BMRC), and Grant No. NSTB EMT/00/022 (CADRA), both of the Agency for Science, Technology, and Research (A*STAR) and the Ministry of Education in Singapore.

References

1. Coronary Artery Disease Overview. http://imaginis.com/heart-disease/cad_ov.asp, accessed on February 19, 2005.
2. What is Coronary Artery Disease? http://www.nhlbi.nih.gov/health/dci/Diseases/Cad/CAD_WhatIs.html, accessed on February 19, 2005.
3. Cooper, G. F., Herskovits, E.: A Bayesian method for the induction of probablistic networks from data. Machine Learning. **9** (1992) 309–347.

4. Friedman, N.: The Bayesian structural EM algorithm. Proceedings of the 14th Conference on Uncertainty in Artificial Intellegience. (1998) 129–138.
5. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning **29** (1997) 131–163.
6. Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning probabilistic relational models. IJCAI '99. Proceedings of the 16th International Joint Conference on Artificial Intelligence. (1999). 1300–1309.
7. Heckerman, D., Geiger, D., Chickering, D. M.: Learning Bayesian Networks the combination of knowledge and statistical data. Machine Learning. **20** (1995) 197–243.
8. Heng, C. K.: Candidate genes for Coronary Artery Disease. Ph.D. Thesis. Department of Paediatrics: National University of Singapore. (1996)
9. Lapuerta, P., Azen, S. P., LaBree L.: Use of neural networks in predicting the risk of Coronary Artery Disease. Computers and Biomedical Research. **28** (1995) 38–52.
10. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics. **6** (1978) 461–464.
11. Tham, C. K., Heng, C. K., Chin, W. C.: Predicting risk of Coronary Artery Disease from DNA microarray based genotyping using Neural Networks and other statistical analysis tool. Journal of Bioinformatics and Computational Biology. **1** (2003) 521–539.
12. Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger A. M., Silbershatz, H., Kannel, W. B.: Prediction of Coronary Artery Disease using risk factor categories. Circulation. (1998) 1837–1847.
13. Witten, I. H., Frank, E.: Data mining: practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann. (1999)