Probabilistic Models for Cardiovascular Events

Kate Sanders, Robert Long +, Nandini Ramanan *, Sriraam Natarajan

Hendrix University; + DePauw University; * University of Texas at Dallas

Abstract

Coronary Heart Disease (CHD) is a leading cause of death in the United States. There are many risk factors of CHD, and these risk factors have complex interactions over the course of decades. Coronary Artery Calcification (CAC) is one indicator of CHD. We used score-based, constraint-based, and local discovery algorithms to learn Bayesian Network (BN) structures for each year of observed data (0, 5, 7, 10, 15, 20) in the Coronary Artery Risk Development in Young Adults (CARDIA) study. These networks model the influence of various clinical and non-clinical risk factors on CAC levels. After comparing the BNs, we selected the most accurate model of the data. Models such as the one selected could enable physicians to construct a more individualized treatment approach for young adults, reducing their risk of CHD later in life.

1 Introduction

According to the American Heart Association, an estimated 16.5 million Americans over the age of 20 suffer from Coronary Heart Disease (CHD). Approximately 1 in 7 deaths that occur in the United States are a result of CHD. The most common and deadly cardiac event associated with CHD is a Myocardial Infarction (MI), often referred to as a heart attack. Each year, approximately 790,000 MIs occur, which means, on average, an American has an MI every 40 seconds (Benjamin et al. 2017).

Coronary Artery Calcification (CAC) is predictive of major cardiac events MI and death from CHD. Detrano et al. found that doubling CAC levels caused an approximately 25% increase in the probability of a major cardiac event occurring, a correlation which held true across all races (Detrano et al. 2008).

To model causes of CAC, we made Bayesian Networks using CARDIA data. We constructed multiple BNs for each year, displaying the relationships between the given attributes. We then implemented three different score based algorithms and three structure learning algorithms on our BNs, enabling us to make our predictions more accurate. The score based algorithms we used are Bayesian information criterion (BIC), Akaike information criterion (AIC), and Bayesian Dirichlet equivalence(BDe). The structure learning algorithms we used are Semi-Interleaved Hiton PC, Chow-Liu, and incremental association. After implementing all six algorithms, we constructed twelve more Bayesian networks: for years 15 and 20 we created a union and intersection separately for both the score based and structure learning algorithms, and a union and intersection of all six algorithms together from year 15 and 20.

2 Background

2.1 Introduction to Bayesian Networks

Bayesian Networks (BNs) are probabilistic models often used in Machine Learning. BNs can be used to model real-world data through a system of nodes and edges. Each node represents a feature, such as cholesterol levels or age. Directed edges connect parent nodes to child nodes, qualitatively representing conditional relationships. Each node can have multiple parents and children. In order to make certain assumptions of independence, BNs must be directed, acyclic graphs (DAGs). This means that the direct edges cannot make cycles within the network (Russell and Norvig 1995).

The conditional probability table (CPT) attached to each child node details the quantitative effect of the parents. CPTs are useful for looking up the probability of a feature taking a certain value based on observed features.

2.2 Structure Learning

The structure of a BN is often created by a domain expert, leaving just the parameters of the network to be learned from the data. In our research, we are learning both the structure and the parameters of the BN using the data. Structure learning can be done using search-and-score techniques, constraint-based methods, or a combination of the two (Koski and Noble 2012).

For our purposes, we will focus on search-and-score structure learning. In search-and-score algorithms, many structures are created from the data and scored. The network with the best score is returned as the optimal model for the data.

In this study, we use a Greedy Local Search algorithm called Hill-Climbing (HC). This algorithm starts with an edgeless network. One at a time, a new edge is created and the resulting network is scored. If the score improves, the new edge is kept. This continues until the score no longer improves with new edge additions. At this point, the network is considered to be at the top of a hill.

There are three primary drawbacks of HC algorithms. One is the tendency to get stuck at a local maximum rather than continuing on to the global maximum, the truly optimal network. Another drawback is getting lost in a plateau, where no direction causes a significantly better score. Ridges can also hinder progress; the searcher zig-zags over the ridge while only making slight progress towards the top of the hill. These problems can be minimized by using random restarts throughout the search process to better explore the space (Russell and Norvig 1995).

Structure scoring metrics for HC include Log Likelihood (LL), Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and Likelihood-Equivalence Bayesian Dirichlet (BDe). Each of these is described in detail below.

Log Likelihood is the simplest scoring metric; the top score goes to the structure that fits the data best. The sum is as follows:

$$LL(B|D) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log(\frac{N_{ijk}}{N_{ij}}) \tag{1}$$

where *n* is number of features, q_i is the total possible configurations of the parents of a particular feature, and r_i is the number of states for a particular feature. N_{ijk} represents the total number of times that a feature takes it's k^{th} value and the the feature's parents are in their j^{th} configuration within the given data.

Using LL alone can lead to an excessive number of parameters. This is known as overfitting. An overfitted model is very specific to the data it is trained on and does not generalize well to other data sets. AIC and BIC improve on LL by adding a penalizing term for network complexity (Koski and Noble 2012).

AIC was developed by Hirotugu Akaike (Akaike 1974) in the early 1970s. The score can be calculated using the following equation:

$$AIC(B|D) = LL(B|D) - |B| \tag{2}$$

where the DAG along which the factorization made is represented by *B* and the data is represented by *D*. The penalizing term is $|B|$, which represents the network complexity in terms of the number of parameters in *B*.

BIC was created in 1978 by Gideon Schwarz (Schwarz 1978) as an alteration on AIC. It can be calculated by using:

$$BIC(B|D) = LL(B|D) - \frac{1}{2}\log(N)|B| \tag{3}$$

The sample size of *B* is represented by *N*. The penalizing factor for BIC is greater than that of AIC.

Unlike BIC and AIC, BDe uses a Bayesian approach to scoring. This method was developed by Heckerman, Geiger, and Chickering in 1995 (Heckerman, Geiger, and Chickering 1995).

3 Related Works

3.1 Previous Models of CARDIA Data

The development of CAC has been modeled by Dynamic Bayesian Networks (DBNs) using data collected in the Coronary Artery Risk Development in Young Adults (CARDIA) study. This temporal model took into account only non-clinical data to identify how life-style decisions young adults make influence CAC levels later in life (Yang et al. 2015).

CAC level development has also been predicted from CARDIA's measurements of known risk factors such as age, cholesterol, and BMI. This was done using Statistical Relational Learning (SRL) algorithms, specifically Relational Probability Trees (RPTs) and Relational Functional Gradient Boosting (RFGB). The AUC-ROC for RPT was 0.778 ± 0.02. For RFGB the AUC-ROC was 0.819 ± 0.01 (Natarajan et al. 2013).

4 Methods

4.1 Data

The data used was gathered through the Coronary Artery Risk Development in Young Adults (CARDIA) study. This study followed 5115 subjects from 1985-6 until present. Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA served as centers for data collection. Each location recruited participants in a way that ensured an even distribution of sex, race, education level, and age-group (18-25 or 25-30). Data gathered from participants included physical measurements, clinical tests, and an in-depth questionnaire about lifestyle and socioeconomic status. The specifics of the study procedures are detailed elsewhere (Friedman et al. 1988). The breadth of recorded features and the longitudinal nature of this study makes it a good data set for studying the development of heart disease. Having such an in-depth data set enables us to pinpoint risk factors in young adults that

Feature	Divisions
sex	2
race	2
cac	2
hbp	3
smoker	3
age	Quintiles
education	Quintiles
heavy exercise	Quintiles
moderate exercise	Quintiles
bmi	Quintiles
triglicerides	Quintiles
cholesterol	Quintiles
ldl	Quintiles
hdl	Quintiles
glucose	Quintiles
dbp	Quintiles
sbp	Quintiles

Table 1: Features analyzed used in creating the Bayesian Networks along with the data divisions.

have the capability to cause serious cardiovascular complications later on. Providing physicians with the knowledge to install treatments that entail preventative actions in order to improve their patients quality of life in their later years.

The data we analyze comes from years 0 (1985-6), 5 (1990-1), 7 (1992-3), 10 (1995-6), 15 (2000-1), and 20 (2005-6). We explored risk features such as: sex, age, race, education, cholesterol, BMI, HDL level, LDL level, triglycerides, diastolic bp, systolic bp, glucose, exercise, blood pressure, and smoking (the features can be viewed in Table 1.). We split sex, race, and CAC levels into boolean values. Smoker status was placed into 3 categories: non-smoker, does not know if they smoke, and smoker. Also, there are three categories for blood pressure: low, medium, and high. The data from the rest of the features are continuous. We discretized the data to make it easier to produce accurate results while implementing it into our BNs. To do this we cut the data into Quintiles, which means the data was split on the following percentiles based on frequency: 0,0.2, 0.4, 0.8, and 1.0.

In the data cleaning process, we found that it would be most beneficial to exclude some of the subjects from our analysis: There were two reasons for doing this.1) If the participant's CAC levels were unobserved in year 20. 2) If one or more of a participant's features were not recorded for any year of the study. Otherwise, missing data was replaced by the mean of that patient's data from every observed year.

4.2 Creating Bayesian Networks

For BN structure learning from the data, we used the bnlearn package in R (Scutari 2009; **?**). The only ordering we used was defining that the sex, race, and age nodes have no parents.

We used four algorithms for learning the structure of the BN. Hill-climbing was the score-based algorithm used. For both of these implementations, we used the optimized implementation to decrease the number of repeated tests within the learning process (Daly and Shen 2007). The score metrics used in the hill-climbing algorithm were AIC (Akaike 1974), BIC (Schwarz 1978), and BDe (Heckerman, Geiger, and Chickering 1995).

We also implemented three constraint based algorithms for learning the structure of the BN. The three constraint based algorithms we used were Incremental Association, Chow Liu, and Semi-Interleaved Hiton-PC. Incremental Association and Semi-interleaved Hiton PC algorithms learn the equivalence class of a directed acyclic graph (DAG) from the data set that is presented. They use conditional independence tests to determine the Markov blankets of the attributes, which are utilized to calculate the structure of the BN. Chow Liu discovers simple tree structures from a data set using pairwise mutual information coefficients.(Scutari 2017)

For every year data was collected (0, 5, 7, 10, 15, 20), we created a BN modeling the features in Table 1 using each of the above algorithms. We printed out all of these networks and compared the dependency connections. We found that the constraint based structure learning algorithms created models that were very sparse. Given the sparsity of the models, we were not able to gather a lot of useful results. Thus, we focused on the models that Incorporated the hill-climbing structure learning algorithm, using the AIC, BIC, and BDe scoring metrics. We placed a heavy focused on the models from year 20 that used AIC, BIC, and BDe, because this is when people had observable high CAC levels. With our three models from year 20 that use the AIC, BIC, and BDe scoring metrics we created a intersection model, which can be viewed 1.

The intersection model shows both some interesting and expected relationships. It is presumed that the blood pressure and cholesterol measurements independently, influence each other. For example, cholesterol influencing LDL and HDL influencing LDL, or SBP influencing DBP which influences HBP. Alongside, a person who can perform heavy (intense) exercise can also perform moderate exercise. The interesting correlations that can be observed from this model are: sex and BMI influencing glucose, sex and glucose influencing CAC, race influencing education, and both race and education influencing if someone smokes.

5 Findings

- BN structures were learned using Hill-Climbing, Grow and Shrink, Incremental Association, and Semi-Interleaved HITON-PC. For the hill-climbing algorithm, three different scoring metrics were used: BDe, BIC,

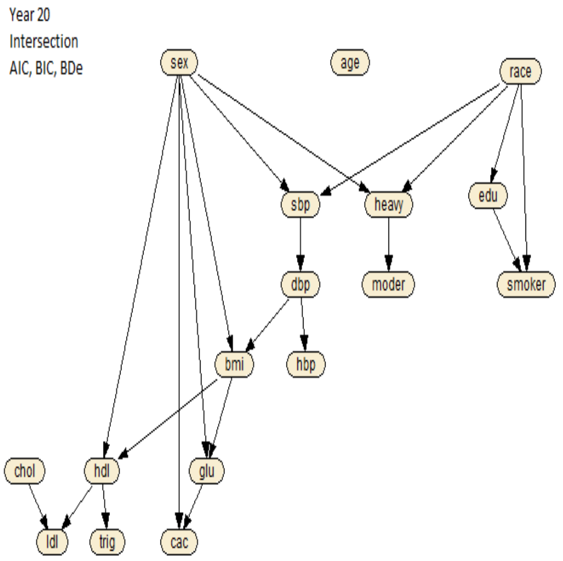

Figure 1: Year 20 Intersection Model of AIC, BIC, BDe

and ML. After training our data we applied each of the scoring functions on our network and found that (insert best method) was the most accurate way to score the data.Thus, we used the results from this network to make our predictions.

- Insert the photos of the best network
- Provide a Graph of the AUC-ROC curve which illustrates a comparison of the different classifiers.

As previously stated, we discovered some interesting relationships between the risk factors. These interesting correlations are sex and glucose influencing CAC, as well as, sex and BMI influencing glucose. We observed that for both sexes, the probability of CAC being present increases as a person's glucose levels rise. However, for each glucose level, men are approximately twice as likely to have CAC than women. Since glucose levels play a large role in CAC presence we also wanted to analyze the risk factors that influence glucose. From this, we saw that BMI and sex influences glucose, such that, those with a low BMI had low glucose levels and those with a high BMI had high glucose levels. Although, at any BMI, men have higher glucose levels, which directly influences CAC levels.

We also found some fascinating relationships between some socioeconomic factors that weren't directly related to CAC. The socioeconomic correlations we discovered are, race influencing education and both race and education having an influence on smoking. In both races, a majority of people completed twelve years or more of education, which means graduating high school. However, the black population had a high probability of receiving twelve years of education or less and a low probability of attending graduate school. This was opposite for the white population, which had a high probability of attending graduate school and a low probability of receiving twelve years of education or less.

Also, for both races, the probability of a person being a non-smoker increases dramatically as higher education is completed. The probability of an individual being a smoker is highest when they have completed twelve years of education or less, and lowest when the individual pursues a graduate degree. We also saw how race played a role as well, with the black population being generally more likely to smoke across the board.

6 Discussions

We have created a network for predicting CAC levels with an [insert precision/recall]. This network could contribute to early identification of patients at high risk for CHD based on longitudinal EHRs that take into account both clinical and non-clinical information. By highlighting the early causes of CHD, doctors can also better advise young patients on precautionary measures for decreasing the long-term risk of CHD. This could contribute to an overall decrease in deaths attributed to CHD.

Future work will focus on testing this network's predictive abilities on different data sets and expanding the network to encompass more risk factors.

7 Results and Discussions

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6):716–723.

Benjamin, E. J.; Go, A. S.; Arnett, D. K.; Blaha, M. J.; Cushman, M.; de Ferranti, S.; Despres, J.-P.; Fullerton, H. J.; Howard, V. J.; Huffman, M. D.; Judd, S. E.; Kissela, B. M.; Lackland, D. T.; Lichtman, J. H.; Lisabeth, L. D.; Liu, S.; Mackey, R. H.; Matchar, D. B.; McGuire, D. K.; Mohler, E. R.; Moy, C. S.; Muntner, P.; Mussolino, M. E.; Nasir, K.; Neumar, R. W.; Nichol, G.; Palaniappan, L.; Pandey, D. K.; Reeves, M. J.; Rodriguez, C. J.; Sorlie, P. D.; Stein, J.; Towfighi, A.; Turan, T. N.; Virani, S. S.; Willey, J. Z.; Woo, D.; Yeh, R. W.; and Turner, M. B. 2017. *Heart Disease and Stroke Statistics–2017 Update: A Report From the American Heart Association*, volume 131.

Daly, R., and Shen, Q. 2007. Methods to Accelerate the Learning of Bayesian Network Structures.

Detrano, R.; Guerci, A. D.; Carr, J. J.; Bild, D. E.; Burke, G.; Folsom, A. R.; Liu, K.; Shea, S.; Szklo, M.; Bluemke, D. A.; O'Leary, D. H.; Tracy, R.; Watson, K.; Wong, N. D.; and Kronmal, R. A. 2008. Coronary calcium as a predictor of coronary events in four racial or ethnic groups. *The New England journal of medicine* 358(13):1336–45.

Friedman, G. D.; Cutter, G. R.; Donahue, R. P.; Hughes, G. H.; Hulley, S. B.; Jacobs, D. R.; Liu, K.; and Savage, P. J. 1988. Cardia: study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology* 41(11):1105–1116.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20:197–243.

Koski, T. J. T., and Noble, J. M. 2012. A Review of Bayesian Networks and Structure Learning. *MATHEMATICA APPLICANDA* 40(1):53–103.

Natarajan, S.; Kersting, K.; Joshi, S.; and Saldana, S. 2013. Early prediction of coronary artery calcification levels using statistical relational learning. In *IAAI*.

Russell, S. J., and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach*, volume 9.

Schwarz, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6(2):461–464.

Scutari, M. 2009. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* VV(Ii):22.

Scutari, M. 2017. Package bnlearn.

Yang, S.; Kersting, Kristian (TU Dormund, Dortmund, G.; Terry, J. G.; Carr, John Jeffrey (Vanderbilt University, Nashville, T. U.; and Natarajan, Sriraam (Indiana University, Bloomington, U. 2015. Modeling Coronary Artery Calcification Levels from Behavioral Data in a Clinical Study. In *Artificial Intelligence in Medicine (AIM): 15th Conference on Artificial Intelligence in Medicine, AIME 2015*, volume 1. 182–187.