Probabilistic Models for Cardiovascular Events

Kate Sanders, Robert Long[+], Nandini Ramaman[*], and Sriraam Natarajan[*]

Hendrix College; DePauw University[+]; Indiana University[*]

Abstract

Coronary Heart Disease (CHD) is the leading cause of death in the United States. CHD is caused by the complex interactions of multiple risk factors over the course of a lifetime. Thus, studying one risk factor at a time cannot capture the full picture of CHD development. To better understand the most pertinent risk factors for CHD, we created Bayesian Networks that model the influence of 16 clinical and non-clinical measurements on Coronary Artery Calcification, a primary indicator of CHD. The data analyzed came from year 20 of the Coronary Artery Risk Development in Young Adults (CARDIA) study. Network structures were learned using the AIC, BIC, and BDe scoring metrics for the Hill-Climbing algorithm. After comparing the predictive capabilities of the Bayesian Networks, we selected the most accurate model of the data for Knowledge Discovery. Of the risk factors analyzed, sex and glucose levels were the most predictive of CHD.

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), which is more commonly known 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 a strong indication of an individual having of CHD. Detrano et al. found that doubling CAC levels caused around a 25% increase in the probability of a major cardiac event occurring, a correlation which held true across all races (Detrano et al. 2008).

Background

Knowledge Discovery

Data Mining (DM) is the process of applying algorithms to large databases to find interesting patterns. These relationships can then be analyzed and modeled for further understanding. This sort of data exploration is referred to as Knowledge Discovery in Databases (KDD). The focus of KDD is identifying probable hypotheses rather than testing known hypotheses (Brockmann, Hufnagel, and Geisel 2006).

One of the most challenging frontiers for KDD is the medical domain. This is largely because most clinical data sets are large, temporal, and incomplete. Also, medical domain knowledge is required for proper interpretation of the exploratory model (Roddick, Fule, and Graco 2003). Despite these challenges, the rewards of successful KDD are significant. DM has been applied to several medical problems with promising results ***add citations***.

Introduction to Bayesian Networks

Bayesian Networks are currently one of the most promising approaches to knowledge discovery (Brockmann, Hufnagel, and Geisel 2006). A Bayesian Network (BN) can be used to create probabilistic models of real-world data through a system of nodes and edges. Each node represents a feature, such as blood pressure or age. Directed edges connect parent nodes to child nodes, qualitatively representing conditional relationships. Each node can have multiple parents and children (Russell and Norvig 1995).

Conditional independence assumptions are necessary to make learning from large databases tractable. BNs incorporate these independence assumptions by connecting nodes with a directed, acyclic graph (DAG). This means that directed edges cannot make cycles within the network (Brockmann, Hufnagel, and Geisel 2006).

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, even when the observed features are distantly related.

Structure Learning Algorithms

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. In each iteration, an edge is created, reversed, or deleted. The resulting network is then scored. If the score improves, the edge operation is kept. This continues until the score no longer improves with new edge operations. 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.

Information Theoretic Scores

LL is the simplest scoring metric; the top score goes to the BN structure, B that best fits the data, D. LL is given by:

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

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. N_{ij} is the value after k is summed out of N_{ijk}.

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 penalizing term is $|B|$. This represents the network complexity and can be calculated as $\sum_{i=1}^{n}(r_i - 1)q_i$ (Ramanan 2017).

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.

Bayesian Scores

A Bayesian approach can also be taken when scoring networks. Bayesian scores are based of Bayes Rule:

$$p(B|D) = \frac{p(D|B)p(B)}{p(D)} \tag{4}$$

where the posterior, $p(B|D)$ is maximized to find the best network structure given the data. $p(D)$ is the marginal probability of the data, $p(D|B)$ is the likelihood of the given data for a particular structure, and $p(B)$ is a prior that changes given the Bayesian scoring metric used (Ramanan 2017).

BDe is a Bayesian scoring metric developed by Heckerman, Geiger, and Chickering in 1995 (Heckerman, Geiger, and Chickering 1995). After much simplification, BDe can be calculated as:

$$BDe(B|D) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\grave{N}_{ij})}{\Gamma(\grave{N}_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\grave{N}_{ijk}+N_{ijk})}{\Gamma(\grave{N}_{ijk})} \tag{5}$$

where Γ is the Gamma function. Score equivalence is achieved using $\grave{N}$, the equivalent same size structure prior. The parameter prior, $\grave{N}_{ijk}$ when a feature is in it's k^{th} value and the the feature's parents are in their j^{th} configuration within the given data. $\grave{N}_{ij}$ is the value when k is summed out of $\grave{N}_{ijk}$. (Ramanan 2017).

Related Works

Previous Models of CARDIA Data

Dynamic Bayesian Networks (DBNs) have been used to model the temporal links between behavioral and socioeconomic data and CAC. Using only non-clinical CARDIA data revealed how life-style decisions young adults make influence CAC levels later in life (Yang et al. 2015).

The presence of CAC has also been predicted from CARDIA's measurements of known risk factors 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).

Methods

The KDD process involves three main steps. The first step deals with data. The researchers must obtain a suitable database for the domain, clean the data, and reduce the number of variables to be analyzed. This step is covered in the *Data* subsection. Next, DM is performed. This involves choosing which machine learning algorithms to use, then applying them to the data. The *Model Creation* subsection outlines our approach to DM. Lastly, the results of DM should be interpreted, evaluated, and applied. This step is covered in the *Findings* as well as the *Future Work* sections. Each of the above steps are iterative, especially DM (Fayyad, Piatetsky-Shapiro, and Smyth 1996). With each iteration, the utility of the system increases.

Data

The Coronary Artery Risk Development in Young Adults (CARDIA) study gathered the data used in our analysis. 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). This data set is often used for studying the development of heart disease due to the breadth of recorded features and the study's longitudinal nature and high retention rate.

The data we focus on in this analysis comes from the year 20 check-in, which occurred in 2005-6. At this time, the retention rate was 72%. In this check-in, approximately 11% of participants had observed CAC.

The features modeled in our network can be viewed in Table 1. We discretized continuous data into Quintiles, five even groups based on frequency.

Feature	Divisions
Sex	Male, Female
Race	Black, White
CAC	Observed, Unobserved
HBP	No, Yes, Not Sure
Smoker	Never, Previously, Currently
Age	Ordinal
Education	Ordinal
Heavy Exercise	Ordinal
Moderate Exercise	Ordinal
BMI	Ordinal
Triglycerides	Ordinal
Cholesterol	Ordinal
LDL	Ordinal
HDL	Ordinal
Glucose	Ordinal
DBP	Ordinal
SBP	Ordinal

Table 1: Features analyzed along with the data divisions.

A participant was excluded from our study if their CAC was not assessed in year 20 or if at least one of a their features were not recorded in any year of the study. Otherwise, missing data imputed using the mean of that patient's data in every observed year.

Model Creation

For BN structure learning from the data, we used the bnlearn package in R (Scutari 2010). Sex, race, and age were defined as having no parents; other than this, no ordering occurred.

We learned BN structures using constraint based algorithms such as Incremental Association, Chow Liu, and Semi-Interleaved Hiton-PC. However, due to insufficient data, these networks were too sparse for effective knowledge discovery.

The Hill-Climbing algorithm for structure learning yielded more meaningful results. We used the optimized implementation of this algorithm to decrease the number of repeated tests within the learning process (Daly and Shen 2007).

The BIC, AIC, and BDe scoring metrics were used with the Hill-Climbing algorithm to learn three separate models of the year 20 CARDIA data. For knowledge discovery, we focused on the intersection BN, which can be viewed in Figure 1.

We used the same data and the bnlearn package to fit parameters to each of of the BNs. The CPTs for each node allowed us to quantitatively analyze the correlations between features.

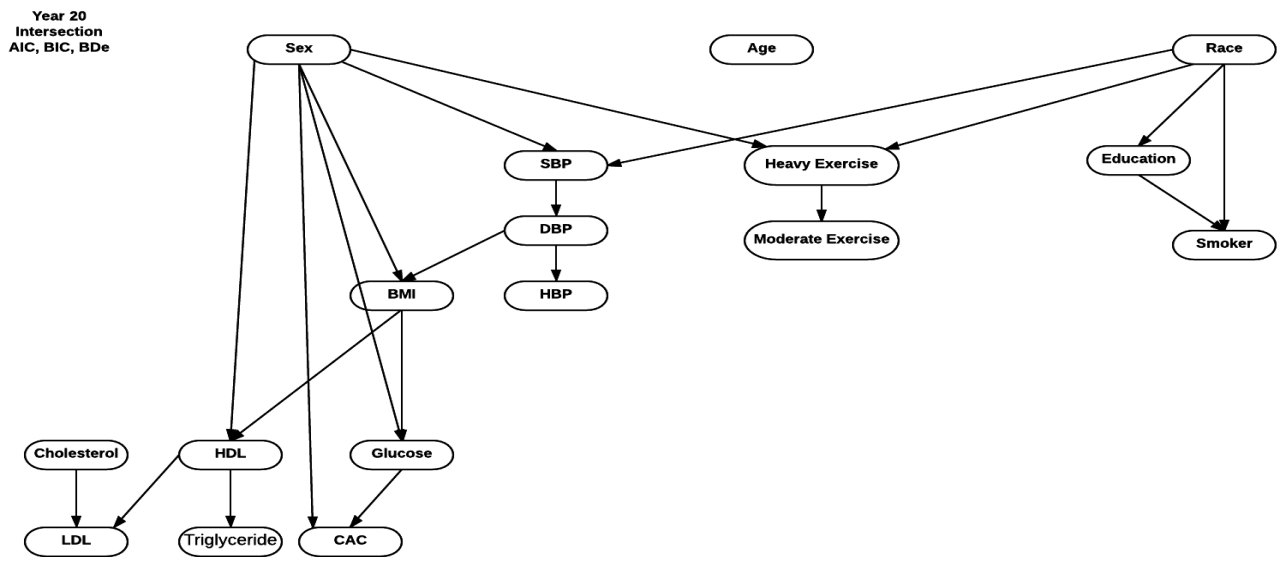

Figure 1: Intersection of models learned using the AIC, BIC, and BDe scoring metrics in the Hill-Climbing algorithm.

Findings

When looking at our intersection network (Figure 1), we first identified three clusters of interdependent measurements to verify the model. The most obvious connection is between the ability to perform moderate and heavy exercise. Unsurprisingly, those capable of performing heavy exercise can also perform moderate exercise. We also saw a monotonic relationship between SBP and DBP, as well as DBP and HBP. This makes sense because an individual with a high systolic blood pressure (SBP) often also has a high diastolic blood pressure (DBP). High blood pressure (HBP) is the diagnosis of individuals with a high SBP and DBP. The cluster of triglycerides, cholesterol, HDL, and LDL was expected as well because all four of these measurements have to do with lipids in the blood stream ***(I can expand on this if needed)***.

Since Figure 1 captured these known connections, we were able to look at more interesting correlations. We first looked at how glucose levels and sex influence the probability of an individual having CAC. For both sexes, there was a monotonic relationship between glucose levels and the probability of CAC. At each glucose level, men were considerably more likely to have CAC than women. It should be noted that, overall, the likelihood of an individual having CAC in year 20 is very low; only approximately 11% of participants had observable CAC at this time.

The risk factors with the most influence on a person's glucose levels were sex and BMI. There was a monotonic relationship between an individual's BMI and glucose levels. As with CAC, men were generally more likely to have high glucose levels.

This model also captured some interesting non-clinical correlations. The most eye-catching connection was between race and education level. In year 20, a majority of participants had pursued post-secondary education (more than 12 years of school). However, African American participants were more likely than Caucasian participants to have a total of 12 years of education or fewer. Caucasian participants were more likely to have pursued a graduate degree (over 16 years of education). This striking difference is evidence of the ongoing effects of racism and segregation in America (Blanchett, Mumford, and Beachum 2005).

A person's race and education also has an influence on whether that person smokes. For both races, the likelihood of an individual smoking decreases as their education level rises. Those with 12 or fewer years of education are much more likely to smoke than those who have pursued a graduate degree. African Americans were more likely to smoke than Caucasians at all education levels. This is likely due to the promotional targeting of minority communities by Big Tobacco (Robinson et al. 1992). Cigarette advertising in minority-specific media pushes menthol cigarettes, which are much more addicting than the non-menthol variety (Cummings, Giovino, and Mendicino 1987).

Future Work

Our findings reveal interesting correlations between both clinical and non-clinical data. These connections warrant further exploration. Particularly, we would like to test the generalizability of these observations on other data sets.

We have three other short term goals for expanding this research. First we plan to incorporate more risk factors in our model, focusing on non-clinical data. We will then create a model using data from all available years of the study. Also, we plan to track the progression of risk factors over the course of many years.

Eventually, we hope to contribute to a system that helps doctors identify the risk of a patient developing CHD based on a variety of temporal factors. Using this system, young adults at risk of CHD could be identified and made aware of precautionary steps for avoiding this deadly disease.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*.

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*.

Blanchett, W. J.; Mumford, V.; and Beachum, F. 2005. Urban school failure and disproportionality in a post-Brown era - Benign neglect of the constitutional rights of students of color. *Remedial and Special Education* 26(2):70–81.

Brockmann, D.; Hufnagel, L.; and Geisel, T. 2006. *Data Mining and Knowledge Discovery Handbook*.

Cummings, K. M.; Giovino, G.; and Mendicino, A. J. 1987. Cigarette advertising and black-white differences in brand preference. *Public Health Reports* 102(6):698–701.

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

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

Fayyad, U.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI magazine* 37–54.

Friedman, G.; Cutter, G.; Donahue, R.; Hughes, G.; Hulley, S.; Jacobs, D.; Liu, K.; and Savage, P. 1988. Cardia: study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*.

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

Koski, T., and Noble, J. 2012. A Review of Bayesian Networks and Structure Learning. *MATHEMATICA APPLICANDA*.

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

Ramanan, N. 2017. *Research on Bayes Net Structure Learning with Complete Data*. Qualifying paper, Indiana University.

Robinson, R. G.; Barry, M.; Bloch, M.; Glantz, S.; Jordan, J.; Murray, K. B.; Popper, E.; Sutton, C.; Tarr-Whelan, K.; Themba, M.; and Younger, S. 1992. Report of the tobacco policy research group on marketing and promotions targeted at african americans, latinos, and women. *Tobacco Control* 1:S24–S30.

Roddick, J.; Fule, P.; and Graco, W. 2003. Exploratory medical knowledge discovery: Experiences and issues. *ACM SIGKDD Explorations Newsletter* 2–7.

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

Schwarz, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics*.

Scutari, M. 2010. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*.

Yang, S.; Kersting, K. (TU Dormund, Dortmund, G.; Terry, J. G.; Carr, J.J. (Vanderbilt University, Nashville, T. U.; and Natarajan, S. (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*.