Modeling Coronary Artery Calcification Levels From Behavioral Data in a Clinical Study

Shuo Yang, Kristian Kersting[+], Greg Terry[*], Jefferey Carr[*], Sriraam Natarajan

Indiana University, USA; [+] TU Dortmund, Germany; [*] Vanderbilt University, USA

Abstract. Cardiovascular disease (CVD) is one of the key causes for death worldwide. We consider the problem of modeling an imaging biomarker, Coronary Artery Calcification (CAC) measured by computed tomography, based on behavioral data. We employ the formalism of Dynamic Bayesian Network (DBN) and learn a DBN from these data. Our learned DBN provides insights about the associations of specific risk factors with CAC levels. Exhaustive empirical results demonstrate that the proposed learning method yields reasonable performance during cross-validation.

1 Introduction

Cardiovascular disease (CVD) is one of the key causes of death worldwide. It is well known that successful and established lifestyle intervention and modification can result in prevention of the development of cardiovascular risk factors. In this work-in-progress, we consider a clinical study, Coronary Artery Risk Development in Young Adults (CARDIA), to model the development of Coronary Artery Calcification (CAC) amounts, a measure of subclinical Coronary Artery Disease [1]. For modeling this risk factor development in adults, we employ the use of temporal-probabilistic models called Dynamic Bayesian Networks (DBNs) [3]. We employ standard optimization scoring metrics [3] – Bayesian Information Criterion (BIC), Bayesian Dirichlet metric (BDe) and mutual information (MI) – for learning these temporal models. We combine the different probabilistic networks resulting from different metrics and evaluate their predictive ability through cross-validation.

One aspect of our work is that in order to learn these DBNs, we consider non-clinical data. We use only basic socio-demographic information and health behaviour information for predicting the incidence of CAC-levels as the individual ages from early to middle adult life. Our results indicate that these behavioral features are reasonably predictive of the occurrence of increased CAC-levels. They allow us to potentially identify potential life-style and behavioral changes that can minimize cardiovascular risks. In addition, the interpretative nature of the learned probabilistic model allows us to easily present the learned models to domain experts (physicians) who could potentially interact with the model and modify/refine the learned model based on their experience/expertise.

2 Background

One of the questions that the CARDIA study tries to address is to identify the risk factors in early life that have influence on the development of clinical CVD in later life. It is a longitudinal population study started in 1985-86

and performed in 4 study centers in the US and includes 7 subsequent evaluations (years 2, 5, 7, 10, 15, 20, 25). It includes various clinical and physical measurements and in-depth questionnaires about sociodemographic background, behavior, psychosocial issues, medical and family history, smoking, diet, exercise and drinking habits. We consider the demographic and socio-economic features to predict the CAC-level as a binary prediction task. Specifically, we consider these features (with their number of categories in the following parentheses) – participant's education level(9), full time(3)/part time(3) work, occupation(8), income(6), marriage status(8), number of children(3), alcohol usage(3), tobacco usage(3) and physical activities(6) during the last year – to model the development of CAC level (High($CAC > 0$)/Low($CAC = 0$)) in each year of study.[1]

3 Proposed Approach

The first issue with this study is that there are several missing values. Preprocessing is required for matching the solutions for missing values across study centers. While the participant retention rate is relatively high (91%, 86%, 81%, 79%, 74%, 72%, and 72%, respectively for each evaluation), there are still at least 10 percent of the data are missing from the records. When a subject is absent from a certain subsequent test, we fill in the missing values using the values from his/her previous measurement. For the missing entries due to other unknown reasons, we treat them as a special class.

Another issue is the evolution of the evaluation measurements and the survey design. Some of the questions related to a certain aspect of the sociodemographic background may be divided into multiple questions or combined into one question in the follow up evaluations. For example, since year 15, the question related to the participant's marriage status has an option as "living with someone in a marriage-like relationship" which is a separate question from year 0 to year 10.

Given these challenges, we employed a purely probabilistic formalism of dynamic Bayesian networks (DBNs) that extend Bayesian networks (BNs) to temporal setting. They employ a factorized representation that decreases the dimension from exponential in the total number of features to exponential in the sizes of parent sets. They also handle the longitudinal data by using a BN fragment to represent the probabilistic transition between adjacent time slots which allows both intra-time-slice and inter-time-slice arcs. Finally, cyclic dependencies in time are allowed. For instance, treatment of a disease in the current time can influence the incidence of the disease in the next time which in turn can influence the treatment in the next time step.

In most literature, the probabilistic influence relationships of the DBN (particularly the temporal influences) are pre-specified and only the parameters are learned. However, since we are interested in determining how the CAC-levels evolve as a function of 10 other risk factors, we instead learn the influences by adapting standard BN structure learning algorithms. The most popular approaches for learning BNs are to employ a greedy local search such as hill climb-

[1] For detailed information about the features, please refer to CARDIA online resource at http://www.cardia.dopm.uab.edu/exam-materials2/data-collection-forms.

ing based on certain decomposable local score functions. We consider three different scoring functions – Bayesian Dirichlet (BDe) scores, Bayesian information criterion (BIC) and Mutual Information Test (MIT).Before discussing the scoring functions, we present the high-level overview of our framework in Figure.1. After preprocessing, we run three hill-climbing algorithms using the three dif-

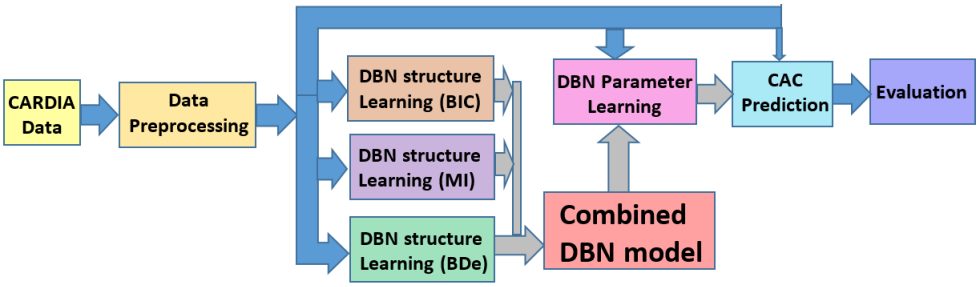

Fig. 1: Flow Chart of the Proposed Model. The blue arrows denote the flow of data and the grey arrows denote the flow of the model.

ferent scoring metrics. First, we transfer the multi-series dynamic data into a training set by extracting every pair of sequential study measurements of every patient as a training instance. After this step, we got 5114 ($|subjects|$) $*5$ ($|paired\ time\ slices|$) instances in total. Using these training instances, we learn three models using the three metrics. We also combine these models by using the union of all the edges to construct a new unified model[2]. The goal is to introduce more dependencies and evaluate if a more complex model is indeed more accurate. For this new unified model, we learn the parameters and perform 5-fold cross-validation for evaluation.

Returning to the scoring function, the first row of Table.1 presents the general form of the decomposable penalized log-likelihood (DPLL) [4] for a BN $\mathcal{B}$ given the data $\mathcal{D}$. $\mathcal{D}_{il}$ is the instantiation of X_i in data point D_l , and PA_{il} is the instantiation of X_i's parent nodes in D_l. So the general form of DPLL is the sum of individual variable scores which equal to the loglikelihood of the data given the local structure minus a penalty term for the local structure. Note that BIC and BDe mainly differ in the penalty term. The penalty term for BIC is presented in the second row where q_i is the number of possible values of PA_i, r_i is the number of possible values for X_i and N is number of examples (5114×5). Hence the BIC penalty is linear in the number of independent parameters and logarthmic in the number of instances. BDe penalty is presented in third row where $\mathcal{D}_{ijk}$ is the number of times $X_i = k$ and $PA_i = j$ in $\mathcal{D}$, and $\alpha_{ij} = \sum_k \alpha_{ijk}$ with $\alpha_{ijk} = \frac{\alpha}{q_i r_i}$ in order to assign equal scores to different Bayesian network structures that encode the same independence assumptions. Ignoring the details, the key is that the complexity of BIC score is independent of the data distribution, and only depend on the arity of random variables and the arcs among them while the BDe score is dependent on the data and controlled by the hyperparameters α_{ijk}. For

[2] When the combination induces intra-slice cycles, we randomly remove one edge.

more details, we refer to [3]. Instead of calculating the log-likelihood, MIT score (Table.1 last row) uses mutual information to evaluate the goodness-of-fit [5]. $I(X_i, PA_i)$ is the mutual information between X_i and its parents. $\chi_{\alpha, l_{i\sigma_i(j)}}$ is chi-square distribution at significance level $1-\alpha$. We refer to [5] for more details.

Table 1: The different scoring functions.

$DPLL(\mathcal{B}, \mathcal{D})$	$DPLL(\mathcal{B}, \mathcal{D}) = \sum_i^n [\sum_l^N \log P(\mathcal{D}_{il} \mid PA_{il}) - Penalty(X_i, \mathcal{B}, \mathcal{D})]$
	$Penalty_{BIC}(X_i, \mathcal{B}, \mathcal{D}) = \frac{q_i(r_i-1)}{2} \log N$
	$Penalty_{BDe}(X_i, \mathcal{B}, \mathcal{D}) = \sum_j^{q_i} \sum_k^{r_i} \log \frac{P(\mathcal{D}_{ijk} \mid \mathcal{D}_{ij})}{P(\mathcal{D}_{ijk} \mid \mathcal{D}_{ij}, \alpha_{ij})}$
$DPMI(\mathcal{B}, \mathcal{D})$	$S_{MIT}(\mathcal{B}, \mathcal{D}) = \sum_{i, PA_i \neq \emptyset} 2N * I(X_i, PA_i) - \sum_{i, PA_i \neq \emptyset} \sum_j^{q_i} \chi_{\alpha, l_{i\sigma_i(j)}}$

4 Experiments

For the DBN DPLL-structure learning, we extended the BDAGL package of Murphy et al. [2] to allow learning from multi-series dynamic data and to support learning with BIC score function. We also adapted DPMI-structure learning by exploiting the GlobalMIT package which was used to model multi-series data from gene expression [5]. The learned DBN structure is shown in Figure.2. Note that all three score metrics learned the self-link for every variable. This shows that many socio-demographic factors are influenced by previous behavior(5 years backward). Observe that both BIC and BDe learned the inter-slice dependency between "Smoke" and "CAC" level. Both BIC and MI returned a temporal correlation between "Exercise" and "CAC", which indicates that previous health behaviors have strong influence on the risk of CAC in current time.

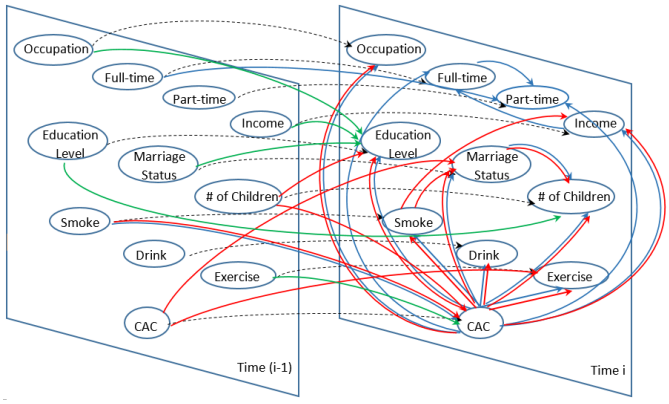

Fig. 2: Combined DBN model. The blue arcs are learned by DPLL-BDe; reds by DPLL-BIC; greens by DPMI; black dash lines are self-links are learned by all three.

Then we combined the structure from the different learning approaches into a comprehensive model for learning parameters. We applied this DBN to test data and predict the CAC score based on the variables in the current and previous time steps. As CAC score from previous time-steps is highly predictive of future values, we hid the CAC-scores in the test set for fair evaluation.

We calculated the accuracy, AUC-ROC as well as F measure[3] to evaluate the independent and mixed models learned by different score functions. The results

[3] Accuracy = (TP+TN)/(P+N); F = 2TP/(2TP+FP+FN)

are shown in Table.2. As the table shows, the model learned by BDe score has the best performance while the MI the worst. We also performed t-test on the five folds results, which shows the BDe is significantly better than MI with P-values at 0.0014(AUC), 0.0247(Accuracy) and 5.7784×10^{-6}(F). In order to rule out the possibility that the better performance of BDe is resulted from the non-temporal information which MI does not have, we also experimented on BNs with intra-slice arcs only as well as DBNs with inter-slice arcs only. And the results showed that the difference between them is not statistically significant at 5% significance level, in other word, the temporal information is as important as the non-temporal information in predicting the CAC level. Compared to BDe alone, the combined model has deteriorated performance. This is probably because the combined model has more arcs which exponentially increases the parameter space and the limited amount of training data cannot guarantee the accurate training for such high dimension model (possibly overfitting).

Table 2: Model Evaluation Results.

	MI	BIC	BDe	MI+BIC	MI+BDe	BIC+BDe	MI+BIC+BDe	Inter-s	Intra-s
Accuracy	0.5774	0.6558	**0.6805**	0.6482	0.6715	0.6701	0.6600	0.5774	0.5853
AUC-ROC	0.4870	0.6979	**0.7139**	0.6473	0.6809	0.7092	0.6581	0.6013	0.6489
F measure	0	0.5417	**0.6144**	0.4640	0.5632	0.5880	0.5244	0.0005	0.0501

Future Work: While our work generatively models all the variables across the different years, extending this work to predict CAC-levels discriminatively remains an interesting direction. Considering more risk factors and more sophisticated algorithms which allows learning temporal influence jumping through multiple time slices is another direction. The end goal is the development of interventions that can reduce the risk of CVDs in young adults.

Acknowledgments: The Coronary Artery Risk Development in Young Adults Study (CARDIA) is supported by contracts HHSN268201300025C, HHSN268201300026C, HHSN268201300027C, HHSN268201300028C, HHSN268201300029C, and HHSN268200900041C from the National Heart, Lung, and Blood Institute (NHLBI), the Intramural Research Program of the National Institute on Aging (NIA), and an intra-agency agreement between NIA and NHLBI (AG0005). SY and SN gratefully acknowledge National Science Foundation grant no. IIS-1343940.

References

1. Detrano, R., Guerci, A.D., Carr, J.J., et al.: Coronary calcium as a predictor of coronary events in four racial or ethnic groups. New England Journal of Medicine 358(13), 1336–1345 (2008)
2. Eaton, D., Murphy, K.: Bayesian structure learning using dynamic programming and MCMC. In: UAI (2007)
3. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
4. Liu, Z., Malone, B.M., Yuan, C.: Empirical evaluation of scoring functions for bayesian network model selection. BMC Bioinformatics 13(S-15), S14 (2012)
5. Vinh, N.X., Chetty, M., Coppel, R.L., et al.: Globalmit: learning globally optimal dynamic bayesian network with the mutual information test criterion. Bioinformatics 27(19), 2765–2766 (2011)