Discriminative Boosted Bayes Networks for Learning Multiple Cardiovascular Procedures

Nandini Ramanan
Proactive Health Informatics
Indiana University
Bloomington, Indiana, USA
nramanan@umail.iu.edu

Shuo Yang
Proactive Health Informatics
Indiana University
Bloomington, Indiana, USA
shuoyang@umail.iu.edu

Shaun Grannis
Regenstrief Institute, Inc
USA
sgrannis@regenstrief.org

Sriraam Natarajan
School of Informatics and Computing
Indiana University
Bloomington, Indiana, USA
natarasr@indiana.edu

***Abstract*—We consider the problem of predicting three procedures, viz, EKG, Angioplasty and Valve Replacement procedures jointly from Electronic Health Records (EHR) and develop a discriminative boosted Bayesian network algorithm. Differences between our proposed approach and standard Bayes Net structure learners are (1) we do not assume that the number of features (observations) are uniform across training examples and (2) our method explicitly handles the precision-recall tradeoff. Our empirical evaluations on a real EHR data demonstrates the superiority of this proposed approach to learning these procedures individually.**

I. Introduction

Coronary heart disease (CHD) is a major cause of death worldwide. In the U.S. CHD is responsible for approximately 1 in every 6 deaths with a coronary event occurring every 25 seconds and about 1 death every minute based on data current to 2007. Effectively predicting whether a particular cardiovascular procedure would be performed on a patient when she/he is admitted to the hospital can help in several tasks including resource allocation, treatment planning and potentially provide the physician with valuable resources needed for making informed decisions. We consider the problem of modeling multiple cardiovascular procedures jointly. Specifically, we consider three of the most common procedures: electrocardiogram (EKG), angioplasty and valve replacement procedures. For example, consider a patient who enters a hospital. Our aim is to build a system that can predict if these procedures are going to be performed on the patient based on his/her clinical measurements along with behavioral data till the admission. Our hypothesis, that we verify empirically, is that joint modeling of these procedures is more effective than modeling each of them separately. Yet another important advantage of predicting these procedures is that it will enable the prediction of future medical costs for a patient and the hospital leading to a better allocation of monetary and hospital resources towards the patient treatment.

In the fields of artificial intelligence and machine learning, joint modeling of events is typically performed using the machinery of probabilistic models, specifically Bayes Networks (BN) [4]. A typical BN consists of two components - a graphical structure that captures the qualitative dependencies between variables/elements in the domain and conditional distributions (or prior distributions) that capture the quantitative dependencies among these elements. The structure learning task of a BN involves learning the parameters in each iteration which could potentially involve probabilistic inference (in the case of hidden data). Given that inference is NP-hard, the problem of structure learning is computationally intensive. Consequently, several different assumptions are made in many learning problems - some methods assume that the order of the variables is known and fixed thus avoiding the complex step of acyclicity checking. The more recent work [6] proposed a decomposable structure learning that approximates the conditional likelihood by factoring it. Inspired by this, we propose a discriminative structure learning algorithm where the loglikelihood is factored by the individual conditional likelihood of the target variables. Specifically, we assume the presence of two types of variables - observed variables/features (in our case, clinical measurements and the behavioral data) and target/modeling variables (in our case, the cardiovascular procedures). Given these separate set of variables, we proceed to learn the conditional distributions of the target variables. To avoid cyclic dependencies in the model, we assume the presence of an ordering among the target variables. This is not a critical assumption as typically, the number of target variables is small enough to search through the space of all possible orderings efficiently. For learning each conditional distribution, we employ the recently successful gradient-boosting methods. The advantage of these methods is that they can simultaneously learn the qualitative and quantitative dependencies in the conditional distribution. Finally, given that our data is highly imbalanced, i.e., the number of patients on whom a certain procedure is performed is smaller than the number of patients who did not have that procedure, we employ a cost-based loglikelihood function that allows us to balance the precision vs. recall (sensitivity vs. specificity) trade-off in a principled manner.

II. Background

A. Discriminative Bayesian Network Learning

Bayesian network (BN) is a probabilistic graphical model which uses nodes to represent random variables, edges for conditional dependencies and conditional probability distributions for the strength of stochastic correlations. There are two different schemes to approach the structure learning of

BNs: generative and discriminative methods. Generative BN learning models the joint probability distributions of the input features and target variable through maximizing the joint likelihood. Discriminative BN Learning directly optimizes the posterior conditional probability of the target variable given the optimal parent set which is captured during the learning process. Discriminative learning approaches are more simple and accurate on predicting compared with generative learning approaches which need to estimate the joint distribution. In medicine, the most common task is to predict certain medical events (diagnosis of diseases, the performance of treatments, etc.) by considering all other information from the patients' electronic health records, such as the lab measurements, demographical attributes, genomic factors, etc. Hence, it is not surprising that discriminative BN learning gets more and more attention in bioinformatics and medical fields [2].

B. Learning DBNs with Functional Gradient Boosting

Functional Gradient Boosting (FGB) is proposed by [9] where gradients are computed over a functional representation of the target distribution. FGB represents the conditional probability distribution of the target variable as a sigmoid over a (potentially non-parametric) function ψ.

$$P(x|Parents(x)) = \frac{\exp \psi(x; parents(x)))}{1 + \exp \psi(x; parents(x)))}$$

where $Parents(x)$ indicates the parent set of the target variable x. It then calculates the gradients for each example of the target variable by computing the functional gradient of the pseudo-loglikelihood objective function. The gradient ($\Delta(x_i)$) for an example x_i is given by $I(x_i = true) - P(x_i = true|parents(x_i))$ where I is the indicator function which returns 1 for positive and 0 for negative examples. These gradients correspond to the difference between the true label and predicted probability of an example. The key insight to FGB is that the gradients are not summed over all the examples, but instead, they are computed for each example and this gradient becomes a weight for that example ($\Delta(x_i)$). Then, a regression tree is learned to fit this regression dataset, which is then added to the model. The sum of the regression values returned by the sequence of trees after m iterations corresponds to the sum of m gradient steps.

III. Joint Learning of CVD Procedures

A standard methodology for learning a distribution over multiple events/targets is the use of Bayes Networks [4]. Typically learning a Bayes Net consists of searching for valid structures and scoring them (specifically for score based methods). To score a network, one has to first estimate the parameters of the updated network and then compute the score. Most scores such as BIC, AIC, BDe etc. consist of two parts - the likelihood of the data (the fit of the model) and a penalty term for model complexity (regularization). Our approach to learning Bayes Nets stems from various observations on this methodology: first, in current Bayes Net learning methods, the search and score are performed sequentially and repeatedly. Our method is capable of simultaneously learning the structure and parameters of the network (i.e., parents of the current target node) since we consider discriminative learning of these models. Second, the use of FGB (specifically, the number of trees) allows us to not consider the penalty term as regularization is taken care of automatically. Third, each conditional distribution in a standard Bayesian network can be potentially represented as a tree to model context-specific independence. Our proposed work can be seen as replacing each single large tree of a conditional distribution with a set of small trees learned in a sequentially boosted manner. Finally, the acyclicity condition is realized implicitly in our discriminative learning method since we assume an order over the target variables as provided by the domain expert. Even in the cases where this order is not provided, checking acyclicity is easier due to a small number of target variables while in general, this is exponential in the size of the feature set. In our experimental setup, the algorithm performs a greedy search over the space of structures i.e., it picks the structure with the maximum likelihood score to obtain the best order as EKG, $Angioplasty$ and $ValveReplacement$. Note that the complexity penalty associated with typical Bayes Net structure learning methods can be ignored as the regularization comes from the boosted trees (depth and number). We use $\mathbf{F}$ to denote the input features such as blood pressure, cholesterol level, triglycerides, smoking status etc. and use T_i to indicating the targets (in our case, cardiovascular procedures - EKG, $Angioplasty$ and $ValveReplacement$) whose conditional probability distributions are being estimated. According to the chain rule of Bayes Networks, the maximum likelihood estimation (MLE) task of the joint learning can be reduced to [1]:

$$\begin{aligned} P(\mathbf{T}, \mathbf{F}) &= P(\mathbf{T}|\mathbf{F}) \cdot LL(\mathbf{F}) \\ &\propto P(\mathbf{T}|\mathbf{F}, \mathbf{T}) = \Pi_i P(T_i|\mathbf{T}_{1;i-1}, \mathbf{F}) \end{aligned} \tag{1}$$

where T_i is the i^{th} target and the target variables $T_{1:i-1}$ is the first $i-1$ target variables according to the order. Given this justification, now, the goal is to then estimate each conditional distribution $P(T_i|\mathbf{T}_{1;i-1}, \mathbf{F})$ separately.

For estimating these distributions, we rely on the machinery of functional-gradient boosting. Specifically, we employ the work of Yang et al. [1] on cost-sensitive FGB. The key idea in this work is to include a cost function that explicitly trades off between false positives and false negatives. For ease of notation, let us consider the current target as y and the set of all the influencing variables (the other targets according to the order and the observed features) as $\mathbf{x}$. $c(y_i, y) = \alpha\, I(y_i = 1 \wedge y = 0) + \beta I(y_i = 0 \wedge y = 1)$, where i is the current example (patient), $I(y_i = 1 \wedge y = 0)$ is 1 for false negatives and $I(y_i = 0 \wedge y = 1)$ is 1 for false positives. Intuitively, $c(y_i, y) = \alpha$ when a positive example (say EKG =1) is misclassified, while $c(y_i, y) = \beta$ when a negative example is misclassified. This

[1]We use bold variables to denote sets. For instance, $\mathbf{F}$ denotes the set of features.

Algorithm 1 Learn DB^2N

Input: $\langle \mathbf{T}, \mathbf{F}, \mathbf{O} \rangle$
▷ **T**: Set of Targets; **F**: Set of Observed Features; **O**: Order of elements in **T**.
Output: BN$\langle \mathbf{N}, \mathbf{E}, \tau \rangle$ ▷ **N**: Set of Nodes, **E**: Set of Edges, τ: Set of multiple trees
for i=1 to $|\mathbf{T}|$ **do**
  $N_i = T_i$ ▷ The current node
  τ_i=SFGBoost(T_i, $\langle \mathbf{F}, \mathbf{T}_{\langle 1:i-1 \rangle} \rangle$) ▷ τ_i models $P(T_i|\mathbf{F}, \mathbf{T}_{\langle 1:i-1 \rangle})$
  E_i=GetFeatures(τ_i)
end for
return $\langle \mathbf{N}, \mathbf{E}, \tau \rangle$

Algorithm 2 SFGBoost : Soft Margin Functional Gradient Boosting

Input: $\langle T_i, \mathbf{F}, \mathbf{T}_{\langle \mathbf{1:i-1} \rangle} \rangle$
$\Psi_0^i :=$ Initial function ▷ i index of the current target
for l=1 to U **do** ▷ Iterate through U gradient steps
  $Tr := GenExamples(i; Data; \Psi_{l-1}^i)$ ▷ Generate example
  $\Delta_l^i = FitRelRegressTree(Tr, \mathbf{F}, \mathbf{T}_{\langle \mathbf{1:i-1} \rangle})$ ▷ Fit trees to the functional gradient
  $\Psi_l^i = \Psi_{l-1}^i + \Delta_l^i$ ▷ Updating the model
end for
return Ψ

leads to a new modified loglikelihood (MLL) function that is now optimized.

$$MLL = \sum_i \log \frac{\exp\left(\psi(\mathbf{x}_i; y_i)\right)}{1 + \exp\left(\psi(\mathbf{x}_i; y') + c(y_i, y')\right)} \tag{2}$$

where i is the current example. Recall from the background section that when using the gradient boosting the gradients are computed for each example separately. Hence, we presented the MLL for each example and now, the gradient of this MLL w.r.t $\psi(y_i = 1; \mathbf{x}_i)$ can be shown as:

$$\frac{\partial \log MLL}{\partial \psi(y_i = 1; \mathbf{x}_i)} = I(y_i = 1; \mathbf{x}_i) - \frac{P(y = 1; \mathbf{x}_i) e^{c(y_i, y=1)}}{\sum_{y_i'} [P(y_i'; \mathbf{x}_i) e^{c(y_i, y_i')}]} \tag{3}$$

The gradients of the objective function can be rewritten compactly as:

$$\begin{aligned} \Delta &= I(\hat{y}_i = 1) \; - \; \lambda P(y_i = 1; \mathbf{x}_i) \\ where \quad \lambda &= \frac{e^{c(y_i, y=1)}}{\sum_{y'} [P(y'; \mathbf{x}_i) \; e^{c(y_i, y')}]}. \end{aligned} \tag{4}$$

As $\alpha \to \infty$, which amounts to putting a large positive cost on the false negatives, $\lambda \to 0$ and the gradients ignore the predicted probability as the gradient is pushed closer to 1 ($\Delta \to 1$), indicating a harsher penalty on misclassified positive examples. On the other hand, when $\beta \to -\infty$, the gradients are pushed closer to 0 ($\Delta \to 0$), indicating more tolerance on misclassified negatives. By setting the parameters

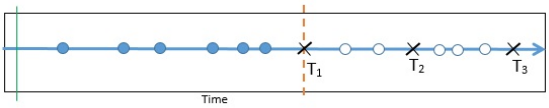

Fig. 1. An example of data extraction. As shown clearly, the data is right censored after the first procedure (shown as T_1). Hence, the goal is to predict the different procedures after observing the risk factors before the first procedure.

$\alpha > 0$ and $\beta < 0$, the different costs of false positive and false negative examples can be incorporated into the learning process. Empirically, we find that any choice of $\alpha \in [0.5, 2]$ and $\beta \in [-2, -10]$ are reasonable choices for efficient learning thus avoiding fine tuning of parameters [1]. Given this gradient, we now proceed to our algorithm. Algorithm 1 presents our *Discriminative Boosted Bayesian Network* (DB^2N) learner. In the outer loop, we consider each target variable in turn and make that variable the current node (N_i) in the Bayes Net. For the current node, we call the soft margin based functional gradient boosting (Algorithm 2) that returns the set of regression trees that model the conditional distribution of this node given its parents. The set of all the variables appearing in this set of regression trees form the parents of the current node (E_i). The process is repeated by going through the order of variables. The final set of nodes with their corresponding edges and the set of trees (for each conditional distribution) are returned.

IV. Experiments

We evaluate the problem of jointly modeling these cardiovascular procedures: EKG, Angioplasty and Valve Replacement, using our DB^2N learning algorithm. We aim to answer the following questions explicitly:

Q1: How does joint modeling of the procedures perform compared to individual learning?
Q2: How does the order of the procedures affect the performance of DB^2N?
Q3: Does DB^2N model the learning task effectively?

We evaluated DB^2N on data extracted from Electronic Health Record from $Regenstrief$ Institute. This dataset has records of 5991 patients, along with their clinical history and some behavioral data, collected between 1973-2015. Our target procedures in the data are *EKG, Angioplasty and Valve Replacement*. We consider several standard risk factors for cardiovascular diseases like HDL levels, smoke status etc.

Data Extraction: When extracting the data, we perform right censoring after the first procedure was performed on the data. This is shown in Figure 1. Since our goal is to predict the procedures jointly after the first procedure is performed, we use the data (risk factors) prior to the first procedure. Note that not all the subjects have all the procedures performed. Hence, it is possible that a single subject could be a positive example for one procedure and negative for another.

We compare our method DB^2N with learning by boosting dependency network where cyclic dependencies are allowed

TABLE I
RESULTS OF RUNNING OUR ALGORITHM AND VARIOUS OTHER CLASSIFIERS FOR PREDICTING ANGIOPLASTY. IT CAN BE OBSERVED THAT DB^2N EXHIBIT THE BEST PERFORMANCE ACROSS ALL THE MEASURES

	Angioplasty				
	Random Forest	Logistic Regression	IPBoost	DNBoost	**DB²N**
AUC PR	0.824	0.751	0.868	0.905	0.887
Precision	0.815	0.690	0.840	0.898	**0.951**
Recall	0.798	0.663	0.766	0.805	**0.909**
F3	0.800	0.666	0.773	0.813	**0.913**
F5	0.799	0.664	0.769	0.808	**0.911**

TABLE II
RESULTS OF RUNNING OUR ALGORITHM AND VARIOUS OTHER CLASSIFIERS FOR PREDICTING EKG. IT CAN BE OBSERVED THAT DB^2N EXHIBIT THE BEST PERFORMANCE ACROSS ALL THE MEASURES

	EKG				
	Random Forest	Logistic Regression	IPBoost	DNBoost	**DB²N**
AUC PR	0.833	0.847	0.861	0.916	**0.919**
Precision	0.783	0.7	0.791	0.857	**0.952**
Recall	0.756	0.76	0.750	0.833	**1**
F3	0.759	0.754	0.754	0.836	**0.995**
F5	0.757	0.758	0.751	0.834	**0.998**

(*DNBoost*). While effective in many domains, they are not interpretable as they approximate the true joint distribution and are possibly cyclic. Our second baseline is the one where each procedure is boosted individually without knowledge of the other procedures (*IPBoost* for individual procedure boost). We additionally compare our classifier with standard, widely used supervised classification methods (Logistic Regression and Random Forest) to predict each procedure individually. Note that these classifiers (Logistic Regression and Random Forest) require fixed length feature vector as their inputs. To this effect, we aggregate each observation using mean function (we explored max, latest etc as aggregators and found mean to be the most informative). We perform 5-fold cross validation and report the area under PR curve along with true positive rate, precision with F3 and F5 scores. Accuracy and log-likelihood are not the useful measures in imbalanced datasets.

The results are summarized in Tables I, II and III for each procedure. As can be seen clearly, DB^2N outperforms all the baselines on almost all the performance measures across all the procedures. This clearly answers $Q1$ and $Q3$ affirmatively. In addition, it achieves very good performance for all metrics on all the procedures. Finally, it is also clear that reasoning about these procedures jointly is more useful than reasoning about them individually. Hence, our method models the medical knowledge of interrelated procedures faithfully. The results in this table for DB^2N employ the following order - *EKG, Angioplasty, valve replacement* - as picked by our algorithm. This answers $Q2$ in that ordering is important when learning a Bayesian network.

TABLE III
RESULTS OF RUNNING OUR ALGORITHM AND VARIOUS OTHER CLASSIFIERS FOR PREDICTING VALVE REPLACEMENT. IT CAN BE OBSERVED THAT DB^2N EXHIBIT THE BEST PERFORMANCE ACROSS ALL THE MEASURES

	Valve Replacement				
	Random Forest	Logistic Regression	IPBoost	DNBoost	**DB²N**
AUC PR	0.748	0.718	0.82	0.866	**0.870**
Precision	0.816	0.684	0.750	0.835	**0.952**
Recall	0.824	0.709	0.773	0.86	0.767
F3	0.823	0.706	0.771	0.857	0.782
F5	0.824	0.708	0.772	0.859	0.772

V. CONCLUSION

We consider the problem of modeling multiple cardiovascular procedures using Bayes Networks. To this effect, we propose an efficient learning algorithm that avoids repeated search and score as is typically done in standard learning methods. Our method effectively learns the parameters and parents of the targets in one step. Experiments on a real EHR demonstrate the effectiveness of this approach in modeling cardiovascular procedures. Extending this algorithm to temporal modeling is an important future direction. We aim to also explore the use of hybrid models in learning discriminative Bayes Networks as our current approach assumes discrete random variables. Finally, more rigorous evaluation of this approach on a larger set of procedures remains an interesting direction.

ACKNOWLEDGEMENT

SN and NR acknowledge the support of Indiana University's Precision Health Initiative. SN and SY gratefully acknowledge National Science Foundation grant no. IIS-1343940. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the PHI or the US government.

REFERENCES

[1] Yang, Shuo, et al. "Learning from imbalanced data in relational domains: A soft margin approach." Data Mining (ICDM), 2014 IEEE International Conference on. IEEE, 2014.
[2] Zhou, Luping, et al. "Learning discriminative Bayesian networks from high-dimensional continuous neuroimaging data." IEEE transactions on pattern analysis and machine intelligence 38.11 (2016): 2269-2283.
[3] Pernkopf, Franz, and Jeff Bilmes. "Discriminative versus generative parameter and structure learning of Bayesian network classifiers." Proceedings of the 22nd international conference on Machine learning. ACM, 2005.
[4] Pearl, Judea. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, 2014.
[5] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." Machine learning 29.2-3 (1997): 131-163.
[6] Carvalho, Alexandra M., et al. "Discriminative learning of Bayesian networks via factorized conditional log-likelihood." Journal of machine learning research 12.Jul (2011): 2181-2210.
[7] Yang, Shuo, and Sriraam Natarajan. "Knowledge intensive learning: Combining qualitative constraints with causal independence for parameter learning in probabilistic models." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2013.
[8] Vinh, Nguyen Xuan, et al. "Polynomial time algorithm for learning globally optimal dynamic Bayesian network." International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, 2011.
[9] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.