

Learning Bayesian Networks using Heart Failure Data

Abstract

Background: Several factors may affect heart failure status of patients. It is important to investigate whether or not the effects are direct. The purpose of this study was learning Bayesian networks that encodes the joint probability distribution for a set of random variables.

Methods: The design was a retrospective cohort study. The target population for this study were heart failure patients who were under follow-up at Asella referral teaching Hospital from February, 2009 to March, 2012. Bayesian Network is used in this paper to examine causal relationships between variables via Directed Acyclic Graph (DAG).

Result: Death of patients can be determined using HIV, hypertension, diabetes, anemia, renal inefficiency and sinus. Hypertension and sinus were found to have direct effects while TB had only indirect effect. Age did not to have an effect.

Conclusion: Anemia, HIV, diabetes mellitus renal inefficiency and sinus directly affect the death of heart failure patient. Death is conditionally independent on TB and age, given all other variables.

Keywords: Bayesian Network, Parameter learning, Structure learning, Causal relationships.

1 Introduction

Graphs provide a comprehensive picture of a problem that makes for a more complete and better balanced understanding than could be derived from tabular or textual forms of presentation. Moreover, graphs can bring out hidden facts and relationships and can stimulate, as well as aid, analytical thinking and investigation.

The main focus of this paper is to describe and drive conditional independence relations existing among random variables through Directed Graphical model called Bayesian networks. Bayesian networks are among the leading technologies to investigate such relations.

A Bayesian network, Cowell et al. (1999), is a graphical model that encodes the joint probability distribution for a set of random variables and a way of finding important relationships between variables. Bayesian Networks provide a powerful technique to understand causal relationships between variables via Directed Acyclic Graph (DAG). Directed graphical models represent probability distributions that can be factored into products of conditional distributions, and have the potential for causal interpretations. The nodes in the graph represent the random variables and missing arrows between the nodes, specify properties of conditional independence between the variables. We refer the reader to Hastie et al. (2008) for detail understanding of Undirected Graphical Models counter part and comprehensive comparison with Directed Graphical Models¹.

A method for learning the parameters and structure of such Bayesian networks has recently been described by Böttcher (2001). Böttcher (2003) described a modern method for learning the parameters and structures of Bayesian networks in deal package of R statistical Software.

Bayesian networks are designed for making decisions in systems with uncertainties Koch (2007). Bayesian networks are therefore suitable for problems where the variables exhibit a complicated dependency structure.

2 Methods

Data Description: The design was a retrospective cohort study, which reviews the patient's card and information sheet. The data of size 263 were obtained from record reviews of all inpatient heart failure patients admitted to Asella Referral Hospital from February, 2009 to March, 2012.

Response Variables: Death status of patients during hospital stay due to heart failure. This status of patient is coded as 1 if the patient died in hospital and 0 if the patient alive.

Independent Variables: The prognostic variables which are expected to be the risk factors of heart failure are categorical and continuous.

¹In an undirected graph, also known as Markov random fields or Markov networks the edges have no directional arrows. The absence of an edge between two vertices indicate that the corresponding random variables are conditionally independent, given the other variables.

Table 1: Independent Variables and their coding

Variables	Category and Coding
Age at the start of treatment	Continues
Renal inefficiency	Yes= 1, No= 0
TB	Positive= 1, Negative= 0
Diabetes mellitus	Positive= 1, Negative= 0
HIV	Reactive= 1, Nonreactive= 0
Anemia	Anemic= 1, Non-anemic= 0
Hypertension	Positive= 1, Negative= 0
Sinus	Positive= 1, Negative= 0

2.1 Bayesian Networks

A Bayesian network is a graphical model that encodes the joint probability distribution for a set of random variables. A graph consists of a set of vertices (nodes), along with a set of edges joining some pairs of the vertices. However, the edges have directional arrows (but no directed cycles) Hastie et al. (2008). The nodes in the graph represent the random variables and missing arrows between the nodes, specify properties of conditional independence between the variables. Bayesian networks are designed for making decisions in systems with uncertainties. Here we perform the analysis using Bayesian networks for discrete and continuous variables in which the joint distribution of all the variables are Conditional Gaussian (CG).

Let $\mathbf{D} = (V, E)$ be a Directed Acyclic Graph (DAG), where V is a finite set of nodes and E is a finite set of directed edges (arrows) between the nodes. The DAG defines the structure of the Bayesian network. To each node $v \in V$ in the graph corresponds a random variable X_v . The set of variables associated with the graph D is then $X = (X_v), v \in V$. Often, we do not distinguish between a variable X_v and the corresponding node v . To each node v with parents $pa(v)$, a local probability distribution, $p(x_v|x_{pa(v)})$ is attached. The set of local probability distributions for all variables in the network is P . A Bayesian network for a set of random variables $\mathbf{X}$ is then the pair (D, P) . The possible lack of directed edges in D encodes conditional independences between the random variables X through the factorization of the joint probability distribution,

$$p(X) = \prod_{v \in V} p(x_v|x_{pa(v)}) \quad (1)$$

Here, Bayesian networks with both discrete and continuous variables are allowed as treated in Bøttcher (2001). So the set of nodes V is given by $V = \Delta \cup \Gamma$, where Δ and Γ are the sets of discrete and continuous nodes, respectively. The corresponding random variables X can then be denoted by $X = (X_v), v \in V = (I, Y) = ((I_\delta), \delta \in \Delta, (Y_\gamma), \gamma \in \Gamma)$, i.e. I and Y used for the sets of discrete and continuous variables, respectively. The set of levels for each discrete variable $\delta \in \Delta$ is denoted as I_δ . To ensure availability of exact local computation methods, we do not allow discrete variables to have continuous parents. The joint probability distribution then factorizes into a discrete part and a mixed part, so

$$p(x) = p(i, y) = \prod_{\delta \in \Delta} p(i_\delta | i_{pa(\delta)}) \prod_{\gamma \in \Gamma} p(y_\gamma | y_{pa(\gamma)}, i_{pa(\gamma)})$$

where $i_{pa(\gamma)}$ and $y_{pa(\gamma)}$ denote observations of the discrete and continuous parents respectively. A method for estimating the parameters and learning the dependency structure of a conditional Gaussian networks with mixed variables is presented in Bøttcher (2001) and implemented in the software package deal in Bøttcher (2003).

2.1.1 Parameter and Structure Learning

To estimate the parameters in the network and to find the structure of the network, a Bayesian approach has been used. So, regarding the parameters, uncertainty about θ is encoded in a prior distribution $p(\theta)$, using data d to update this distribution, i.e. learn the parameters, and here by obtain the posterior distribution $p(\theta | data)$. The section is based on Bøttcher (2001) and Heckerman *et al.* (1995). Consider a situation with one random variable X . Let θ be the parameter to be assessed and Θ be the parameter space and d a random sample of size n from the probability distribution $p(x|\theta)$. We call d our database and $x^c \in d$ a case. Then, according to Bayes' theorem,

$$p(\theta | d) = \frac{p(d|\theta)p(\theta)}{p(d)}, \theta \in \Theta \quad (2)$$

where $p(d|\theta) = \prod_{x^c \in d} p(x^c|\theta)$ is the joint probability distribution of d , also called the likelihood of θ . As prior parameter distributions, the Dirichlet distribution and the Gaussian inverse-Gamma distribution have been used for the discrete variables and for the continuous variables respectively. These distributions are conjugate to observations from the respective distributions and this en-

sure simple calculations of the posterior distributions. Now, to learn the structure of the network, we calculate the posterior probability of the DAG, $p(D|d)$, which from Bayes' theorem is given by

$$p(D|d) = \frac{p(d|D)p(D)}{p(d)},$$

where $p(d|D)$ is the likelihood of D and $p(D)$ is the prior probability of D . As the normalizing constant $p(d)$ does not depend upon structure, another measure, which gives the relative probability, is

$$p(D, d) = p(d|D)p(D),$$

We use the above measure and refer to it as the network score. For simplicity, we choose to let $p(D)$ be the same for all DAGs, so we are only interested in calculating the likelihood $p(d|D)$. So learning the DAG from data, we can in principle first calculate the network scores for all possible DAGs and then select the DAG with the highest network score.

If many DAGs are possible, it is computationally infeasible to calculate the network score for all these DAGs. In this situation it is necessary to use some kind of search strategy to find the DAG with the highest score. In some cases it can be more accurate to average over the possible DAGs for prediction, instead of just selecting a single DAG. So if x is the quantity we are interested in, we can use the weighted average,

$$p(x|d) = \sum_{D \in DAG} p(x|d, D)p(D|d),$$

where DAG is the set of all DAGs and $p(D|d)$ is the weight. Again, if many DAGs are possible, this sum is too hard to compute, so instead, by using a search strategy, we can find a few DAGs with high score and average over these. In order to calculate the network score for a specific DAG D , in a CG network, we need to know the prior probability and the likelihood of the DAG. For simplicity, we could for example choose to let all DAGs be equally likely, then

$$p(D|d) \propto p(d|D),$$

In a CG network, the likelihood of the DAG D is given by

$$p(d|D) = \int_{\theta \in \Theta} p(d|\theta, D) p(\theta|D) d\theta,$$

To evaluate which DAG or possible several DAGs that represent the conditional independences in a Bayesian network well, we want to find the DAG or DAGs with the highest network scores. To calculate these scores, we must specify the local probability distributions and the local prior distributions for the parameters for each network under evaluation. we see in the above equation that it, besides the likelihood of the parameters, also involves the prior distribution over the parameters, $p(\theta|D)$. This means that we for each possible DAG have to specify a prior distribution for the parameters. In Böttcher this method is extended to the mixed case. With this method, the parameter priors for all possible networks can be deduced from one joint parameter prior, called *master prior*. To specify this master prior, we only have to specify a prior Bayesian network, i.e. a prior DAG and a prior probability distribution, together with a measure of how confident we are in the prior network.

2.2 Ethical Considerations

Ethical clearance was obtained from the Hospital.

3 Results and Discussion

For this study, the data of heart failure patients which was taken from Asella Referral Hospital is used and analyzed using Bayesian Networks. The data of size 263 were obtained from record reviews of all in patient heart failure patients admitted to Medical ward from February, 2009 to March, 2012. In this analysis, nine variables are considered each containing 263 observations as presented in table 2.

Table 2: Variables used in this analysis

Node index	Variables
1	Death
2	Age
3	Renal inefficiency
4	TB
5	Diabetes
6	HIV
7	anemia
8	Hypertension
9	Sinus

A Bayesian network is a graphical model that encodes the joint probability distribution for a set of random variables and a way of finding important relationships between variables. Bayesian Networks provide a powerful technique to understand causal relationships between variables via Directed Acyclic Graph (DAG) showing relationships among variables through directed arrow.

Here we consider Bayesian networks with both discrete and continuous random variables. We use the *deal* package of R Statistical Software to analyze the data of heart failure patients obtained from Asella Referral Hospital.

The purpose of analyzing data under this section is to find dependency relations between the variables where the main interest lies in finding out which variables influence the death status of heart failure patients.

Age is continuous variable which may cause the death of heart failure patients. However, in *deal*² continuous parents of discrete nodes are not allowed. Thus, describing such a relation is impossible. A remedial measure is to handle Death as a continuous variable, even though this is clearly not.

Specification of a Bayesian network:

```
> Data<-read.table("mom.dat",header=T) ## invoke Data from working Directory
> attach(Data)
> Data<-data.frame(age,hypertension,HIV,diabetes,TB,anemia,sinus,renalinefficiency)
> Data$death<-as.numeric(Data$death)
> Data$age<-as.numeric(Data$age)
> Data$anemia <-as.factor(Data$anemia)
```

²We recommend to see Bøttcher (2003 and 2003a,b) on this part for more understanding.

```
> Data$diabetes <-as.factor(Data$diabetes)
> Data$hypertension <-as.factor(Data$hypertension)
> Data$sinus <-as.factor(Data$sinus)
> Data$TB <-as.factor(Data$TB)
> Data$HIV <-as.factor(Data$HIV)
> Data$renalineffeciency <-as.factor(Data$renalineffeciency)
```

Hereafter, we are in position to specify a prior Bayesian network. We use the empty DAG as the prior DAG since we have no prior knowledge about specific dependency relations and let the probability distribution of the discrete variables be uniform. The assessment of the probability distribution for the continuous variables is based on data.

```
> library(deal) ## call deal package
> Data.nw<-network(Data) ## specify prior network
> Data.prior <- jointprior(Data.nw) ## create joint prior distribution
Imaginary sample size: 256
# banlist for age and HIV as none of
#the other variables can
influence these variables.
> from1<-c(2,3,4,5,6,7,8,9)
> to1<-rep(1,8)
> from2<-c(1,2,4,5,6,7,8,9)
> to2<-rep(3,8)
> from3<-rep(9,6)
> to3<-c(2,4,5,6,7,8)
> banlist<-matrix(c(from1,from2,from3,to1,to2,to3),ncol=2)
> banlist(Data.nw) <- banlist
```

The ban list is a matrix with two columns. Each row contains the directed edge that is not allowed. The final stage is to learn the parameters in the network and initiate the structural learning using *autosearch()* and *heuristic()*.

```
> Data.nw <- getnetwork(learn(Data.nw,Data,Data.prior))
> Data.search <- autosearch(Data.nw,Data,Data.prior,trace=TRUE)
> Data.heuristic <- heuristic(getnetwork(Data.search),Data,Data.prior,restart=2,
degree=10,trace=TRUE,trylist=gettrylist(Data.search))
```

NB: The banlist forces:

- death to be a leaf node (death can only receive arcs) (in our dataset death is variable number 9)
- Age and HIV cannot receive arcs (in our dataset age is variable number 1 & HIV is variable number 3).

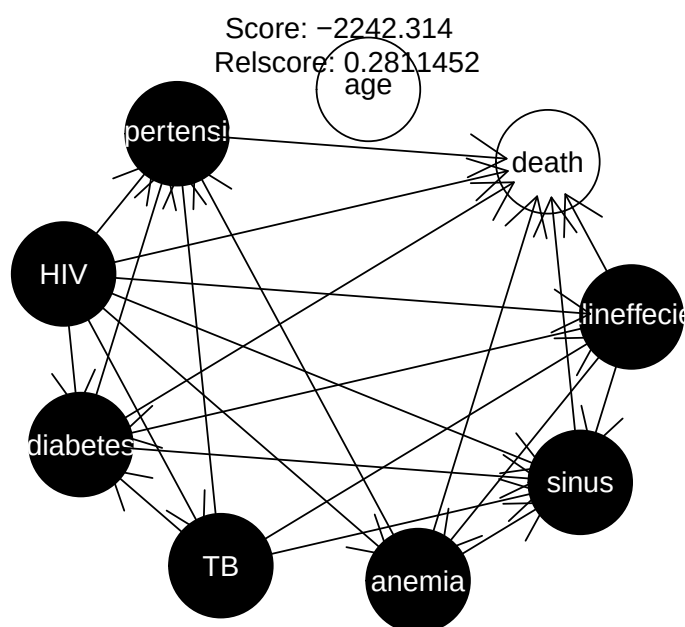


Figure 1: Bayesian Network Plot for Heart failure Data

As can be seen from the Bayesian network plot, Figure1, death depends directly on Human Immune deficiency Virus (HIV), hypertension, diabetes, anemia, renal inefficiency and sinus. Hence, given these variables, death is conditionally independent on TB and age.

Death is conditionally independent on TB, given all other variables. We also see that age is independent from all other variables. How can we explain these findings? TB has an impact on death but only indirectly through complications like HIV, hypertension, diabetes, anemia, renal

inefficiency and sinus. TB does not provoke death directly but only through these events, which absolutely seems very reasonable. Anemia is one of the important variables that can lead to death directly, but also indirectly: through hypertension and sinus. This shows that anemia has somehow three ways to impact death: directly or indirectly, because it increase the possibility to importance of hypertension and sinus.

As a comparison, we see that HIV has a direct effect and also an indirect effect through all other variables. There are many further interesting indirect effects in this network.

Age seems independent from all other variables. This is an interesting finding of the Bayesian network. Age is clearly associated with death in the other studies (like GAM)³, but here, when focus is on conditional independence, we estimated that all other variables have an impact on death, direct or indirect, but age has not. We can understand this in the following way: age is in some sense a surrogate variable for health status, in itself age is not a danger, but through events that become more lively with age, like hypertension. Indirect effects are therefore possible, but not direct ones. Apparently there is not enough support in our data to estimate age as a master node, that from the top regulates all other variables which then point to death.

4 Conclusions

In this paper we described a new powerful technique for analyzing Heart Failure data based on the theory and algorithms for learning Bayesian networks. We explained how to apply these techniques to Heart Failure data. The result of this analysis showed that death of patient can be determined by HIV, hypertension, diabetes, TB, anemia, renal inefficiency and sinus.

³We analyzed these data using Generalized Additive logistic Regression and found all the variables used here including age are associated with death.

5 References

- [1] Heckerman D, Geiger D. and Chickering D (1995) Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20**: 197-243.
- [2] Koch K.R (2007) Introduction to Bayesian Statistics. *Springer, 2nd edition*.
- [3] Cowell R.G , Dawid A.P , Lauritzen S.L, and Spiegelhalter D.J (1999) Probabilistic Networks and Expert Systems. Springer-Verlag.
- [4] Bøttcher S. G (2001) Learning Bayesian Networks with Mixed Variables, *Artificial Intelligence and Statistics 2001*, Morgan Kaufmann, San Francisco, CA, USA, pp. 149-156.
- [5] Bøttcher S. G (2003) Learning Bayesian networks with Mixed Variables. URL <http://www.math.auc.dk/~alma>. Aalborg University.
- [6] Bøttcher S. G and Dethlefsen C (2003a) deal: A Package for Learning Bayesian Networks. *Journal of Statistical Software* **8**(20): 1-40.

- [7] Böttcher S. G. and Dethlefsen C (2003b) Learning Bayesian networks with R, in K. Hornik, F. Leisch and A. Zeileis (eds), *Proceedings of the 3rd international workshop on distributed statistical computing*. ISSN 1609-395X.
- [8] Hastie T, Tibshirani R and Friedman J (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer Series in Statistics*, Second edition, Stanford, California.