Bayesian Network Model for Epidemiological Data (Radiation exposure and circulatory disease risk: Hiroshima and Nagasaki atomic bomb survivor data)

Sagar Baviskar[1], Akash Kabra[2] and Anand Biyani[3]

[1] College of Engineering Pune, Maharashtra, India.

Received: 10 April 2013 Accepted: 1 May 2013 Published: 15 May 2013

Abstract

This documentation describes the implementation of Bayesian Network on Hiroshima Nagasaki atomic bomb survivor data, using ?R? software. Bayesian networks, a state-of-the art representation of probabilistic knowledge by a graphical diagram, has emerged in recent years as essential for pattern recognition and classification in the healthcare field. Unlike some data mining techniques, Bayesian networks allow investigators to combine domain knowledge with statistical data. This tailored discussion presents the basic concepts of Bayesian networks and its use for building a health risk model on Epidemiological data. The main objectives of our study is to find interdependencies between various attributes of data and to determine the threshold value of radiation dosage under which death counts are negligible.

Index terms— bayesian network; data mining; epidemiological data, health risk model, implementation of bayesian network in R.

1 I. INTRODUCTION

ur focus is on identification of the relationships between radiation exposure and its potential risk factors using Bayesian Network, with the emphasis on integrating medical domain knowledge and statistical data analysis.

A Bayesian network is a graphical model that encodes the joint probability distribution for a set of random variables. Here we consider Bayesian networks with mixed variables, i.e. the random variables in a network are both discrete and continuous types.

First, raw data are pre-processed into a format that is acceptable to the learning algorithms of Bayesian networks. Some important considerations are discussed to address the uniqueness of the data and the challenges of the learning.

Second, a Bayesian network is learned from the pre-processed data set by integrating medical domain knowledge and generic learning algorithms. Third, the relationships revealed by the Bayesian network are used for finding the probability of death count.To learn a Bayesian network, the user needs to supply a training data set and represent any prior knowledge available as a Bayesian network. We are implementing The Bayesian Author : Computer Engineering College of Engineering Pune, India. E-mail : Sagarbaviskar91@gmail.com Network in "R" software in one section, we will explain detail implementation of Bayesian Network in R.

This report makes use of data obtained from the Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan. RERF is a private, nonprofit foundation funded by the Japanese Ministry of Health, Labor and Welfare (MHLW) and the U.S. Department of Energy, the latter through the National Academy of Sciences. The conclusions in this report are those of the authors and do not necessarily reflect the scientific judgment of RERF or its funding agencies **??**3] [7].

2 Abbreviations, notations and Acronyms:

Bayesian Network (BN).

3 II. WHAT IS BAYESIAN NETWORK?

Bayesian network is a graphical model where nodes represent random variables (the two terms are used interchangeably in this article) and arrows represent probabilistic dependencies between them.

The graphical structure G = (V; A) of a Bayesian network is a directed acyclic graph (DAG), where V is the node (or vertex) set and A is the arc (or edge) set. The DAG defines a factorization of the joint probability distribution of V = {X1; X2;?;Xn}, often called the global probability distribution, into a set of local probability distributions, one for each variable.

The Bayesian network is a state-of-the art representation of probabilistic knowledge. Bayesian networks represent domain knowledge qualitatively by the use of graphical diagrams with nodes and arrows that represent variables and the relationships among the variables. Quantitatively, the degree of dependencyis expressed by probabilistic terms.

4 III. ADVANTAGES OF BAYESIAN NETWORK AS DATA MINING TOOL

First, Bayesian networks allow investigators to use their domain expert knowledge in the discovery process, while other techniques rely primarily on coded data to extract knowledge. Second, Bayesian network models can be more easily understood than many of the other techniques via the use of nodes and arrows.

5 Year

These represent the variables of interest and the relationships of variables, respectively. Researchers can easily encode domain expert knowledge through the use of these graphical diagrams, and thus more easily understand and interpret the output of the Bayesian network. In addition, Bayesian network algorithms capitalize on this encoded knowledge to increase their efficiency in modeling process and accuracy in its predictive performance. Next, Bayesian networks are flexible in regards to missing information. Bayesian network models can produce relatively accurate prediction even in the situation where complete data are not available. Last, because Bayesian networks can incorporate domain knowledge into statistical data, Bayesian networks are less influenced by small sample size.

A more detailed discussion will enhance the understanding of how Bayesian networks operate and why they are particularly well-suited to the epidemiological data such as Radiation exposure and circulatory disease risk: Hiroshima and Nagasaki atomic bomb survivor data.

IV.

6 BASIC PROBABILISTIC CONCEPTS

Fundamentally, Bayesian networks are designed to, through the complex application of the well-developed probability theory (Bayes rule), obtain probabilities of unknown variables from known probabilistic relationships. To understand Bayesian networks, basic concepts such as the Bayesian probability approach, prior (or unconditional) probability, posterior (or conditional) probability, joint probability distribution, and Bayes rule, need to be discussed.

7 V. BAYESIAN PROBABILITY VS CLASSICAL PROBABILITY

There are differences between Bayesian probability and classical probability. The Bayesian probability of an event is a person degree of belief in that event; the classical probability is the probability that an event will occur. Contrary to classical probability, we do not need repeated trials to measure the Bayesian probability. Thus, Bayesian probability based on personal belief is useful where the probability cannot be measured, even by repeated experiments.

8 VI. PRIOR PROBABILITY

In a situation when no other information (evidence) I available, the probability of an event occurring is a prioror unconditional probability. The commonly used denotation of prior probability is P (A), where the event of A is occurring. Prior probability, P (A), is used only whenno other information is available. Also, denotation, P (^A), can be used to represent the prior probability of an event not occurring. For example, suppose Ineffective Airway Clearance denotes a binary variable whether or not a particular patient admitted in hospital has anursing diagnosis of Ineffective Airway Clearance. The prior probability of Ineffective Airway Clearance may be expressed (estimated) as P (Ineffective Airway Clearance) = 0.15, meaning that without the presence of any other evidence (information), a nurse may assume that a particular patient has a 15% chance of having an Ineffective Airway Clearance. In this example of P (Ineffective Airway Clearance), we can assume that they can have values such as present or absent. Thus, P (Ineffective Airway Clearance) is viewed as P (Ineffective Airway Clearance=present), and P (^Ineffective Airway Clearance) as P (Ineffective Airway Clearance=absent).

A probability term is also used to express random variables with multi-values in the nursing domain. For example, if we are interested in the random variable Cognition of a patient, this variable may have several possible values, such as very good, good, poor, and very poor. We might estimate them based on experience as: P(Cognition=verygood)=0.60; P(Cognition=good)=0.30;P(Cognition=poor)=0.08;and P(Cognition=very poor)=0.02. We can also state all the possible values of the random variable, Cognition, as P (Cognition) = (0.6, 0.3, 0.08, and 0.02), which can be defined as a probability distribution for the random variable Cognition.

9 VII.

10 CONDITIONAL PROBABILITY

As discussed earlier, the probability of an event occurring is expressed as a prior or unconditional probability; once the evidence is obtained, it becomes posterior or conditional probability. Once we have new information B, we can use the conditional probability of A given B instead of P(A), which can be denoted as P(A|B).This means "the probability of A, given B". Suppose P (Ineffective Airway Clearance| Grunting) is estimated to be 0.60. This proposes that if a patient is observed to have a Grunting breathing sound, and no other information is available, and then the probability of the patient having an Ineffective Airway Clearance will be changed from 0.15 to 0.60. That is, without considering the presence of Grunting, the probability ofIn effective Airway Clearance (prior probability) is 0.15; while considering the presence of Grunting, the probability of Ineffective Airway Clearance (posterior probability) becomes 0.60.

11 VIII.

12 JOINT PROBABILITY DISTRIBUTION

The joint probability distribution expresses all the probabilities of all combinations of different values of random variables. As mentioned in the Cognition example, the probability distribution of Cognition is a one dimension alvector of probability for all possible values of a variable. The joint probability distribution is(D D D D) Year 013 2 D

expres sed as an n-dimensional table (n> 1), which is called the joint probability table. The joint probability table consists of the probabilities of all possible events occurring. Table **??** illustrates an example of joint probability distribution with a two-dimensional table of the two variables Pain and Satisfaction with Care in the nursing care domain, in which each variable has three values. Because all events are mutually exclusive, the sum of all the cells is '1' in the joint probability table. This distribution can answer any probabilistic statemen of interest. Adding across a row or column gives the prior probability of a variable; for example, P(Pain=Level I)=0.3 + 0.15 + 0.01 = 0.46. P(Pain=Level I ?Satisfaction with Care=High) can also be obtained which is 0.3.

13 IX. BAYES' RULE

This section demonstrates the details of updating prior probability to conditional (posterior) probability using Bayes_rule. Conditional probabilities can be redefined in Eq. (1),P(A|B) = P(A?B) P(B) (1)

This equation can also be written as:

14 P(A?B)= P(A , B)= P(A|B)P(B)(2) P(A?B)=P(A, B)= P(B|A)P(A)(3).

Based on two equations (Eq. (**??**) and (**??**)), we can induce the equation known as Bayes'rule in Eq. (**??**) (also Bayes' law or Bayes' theorem), by equating the two right hand sides and dividing by P (B)> 0 P(A|B) = P(B|A)P(A)P(B) (4)

Bayes' rule is useful in practice to estimate unknown P (A|B) from three probability terms (i.e., P (B|A), P (A) and P (B)) that nurses may be able to easily estimate in a domain. In a task estimating the probability of Ineffective Airway Clearance, there can be conditional probabilitie son causal relationships as in Fig. **??**: Nurses may want to derive a nursing diagnosis given information by Grunting. A nurse knows that Ineffective Airway Clearance may cause a patient to have a Grunting breathing sound (an estimated 40% of the time). The nurse also knows some unconditional facts: suppose the prior probability of a patient having Ineffective Airway Clearance is 0.15, and the prior probability of any patient having Grunting is 0.10.When a nurse would like toestimateP(Ineffective Airway Clearance|Grunting) which may not be well-known probability, conditional probabilities can be induced based on Bayes' rule in Eq. (**??**). This simple example of Bayes' rule demonstrates how unknown probabilities can be computed from the known.

15 P (Grunting Ineffective Airway

16 X. A TYPICAL BAYESIAN NETWORK

Suppose that there are two events which could cause grass to be wet: either the sprinkler is on or it's raining. Also, suppose that the rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler is usually not turned on). Then the situation can be modeled with a Bayesian network (shown). All three variables have two possible values, T (for true) and F (for false). The model can answer questions like "What is the probability that it is raining, given the grass is wet?" by using the conditional probability formula.

17 XI. ABOUT DATA SET USED

The dataset which is used describe circulatory mortality in the Life Span Study of atomic bomb survivors. It is based on Radiation exposure and circulatory disease risk: Hiroshima and Nagasaki atomic bomb survivor . The data set is a detailed tabulation of person-years, case-counts, and summary data constructed from data on individual survivors. The cohort for analysis includes 86,661 survivors. Data on individual survivors are stratified by city, sex, age at exposure, attained age, calendar time, and dose. Crossclassification variables used to define the table are:1)Name 2) City 3) Sex 4) Agexcat 5) Agecat 6) Ctime. Variables that includes the cell-specific numbers of subjects entering the study:1) Dosecat 2) Subjects 3) PYR 4) Agex 5) Age 6) colon10. Disease death counts variables:1) CVD 2)stroke 3)heartd 4)othcvd 5)concvd 6)constroke 7)conheartd 8)conothcvd ??3][7].

This report makes use of data obtained from the Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan. RERF is a private, nonprofit foundation funded by the Japanese Ministry of Health, Labor and Welfare (MHLW) and the U.S. Department of Energy, the latter through the National Academy of Sciences. The conclusions in this report are those of the authors and do not necessarily reflect the scientific judgment of RERF or its funding agencies ??3][7].XII. IMPLEMENTATION OF BAYESIAN NETWORK IN "R" a) Various BN Algorithms

We are using R software to implement the Bayesian Network for the above dataset.

There are two packages named as 1) Deal and 2) BN Learn using which we can implement the BN.

In Deal package, we have Greedy algorithm to implement the BN for given dataset, and in BN learn we have 1. Constraint based algorithm 2. Score based algorithm In Constraint based algorithm, we can implement the BN by using either Grow-Shrink algorithm or Incremental association markov blanket.

In Score based algorithm package, we have Hill climbing algorithm to build a BN. We will discuss the detailed implementation of BN using Deal package with Greedy algorithm and will just compare the results of rest of the algorithms in the following topics.

18 b) Implementation of BN in R using Deal Package

The data is in a file named as lsscvd10.csv which is a .csv file. So read the data from the csv file into R.

For the implementation we considered following categorical variables (1) City 2) Sex 3) Agexcat 4) Agecat 5) Ctime 6)Dosecatand continuous variables (CVD, STROKE, HEARTD, PYR, COLON10, SUBJECTS) [7]. Now after reading the data, we need to load the data into a data frame which is an acceptable form in R. Now as the first 6 variables are categorical we need to normalize the data by factorizing those variables.

In deal, a Bayesian network is represented as an object of class network. The network object is a list of properties that are added or changed. By default it is set to the empty network (the network without any arrows) [1].

If the option specify graph is set, a point and click graphical interface allows the user to insert and delete arrows until the requested DAG is obtained. Note that discrete nodes are grey and continuous nodes are white.

The primary property of a network is the list of nodes. Each entry in the list is an object of class node representing a node in the graph, which includes information associated with the node. Several methods for the network class operate by applying an appropriate method for one or more nodes in the list of nodes. The nodes appear in the node list in the same order as in the data frame used to create the network object.

The parameters of the joint distribution of the variables in the network are then determined by the function joint prior () with the size of the imaginary data base as optional argument. If the size is not specified, deal sets the size to a reasonably small value [1].

Then comes the most important function which is Learn () function. Using this function, the dataset is learned by R software for finding the relationship between the covariates, i.e, Random variables [1].

To get the best BN the heuristic searching technique is used. The search algorithm is used with restarts which is implemented in the function heuristic (). The initial network is then perturbed according to the parameter degree and the search is performed starting with the perturbed network [1].

After that using fit () function we can compute the probabilities which are desired

19 b) Incremental Algorithm

This algorithm is based on the Markov blanket detection algorithm of the same name, which is based on a two-phase selection scheme (a forward selection followed by an attempt to remove false positives). This algorithm is a variant of Incremental Association which uses speculative stepwise forward selection to reduce the number of conditional independence tests [2].

20 c) Grow Shrink Algorithm

This algorithm is based on the Grow-Shrink Markov Blanket, the first (and simplest) Markov blanket detection algorithm used in a structure learning algorithm. But all the above algorithms are not finding exact relationship between the variables, it means the dependency between variables is unidirectional. So exact inference cannot be drawn using these algorithms. Also these algorithms are not exhaustive.

21 d) Hill Climbing Algorithm

This algorithm finds the optimal network structure in the restricted space. A hill climbing greedy search on the space of the directed graphs. The optimized implementation uses score caching, score decomposability and score quivalence to reduce the number of duplicated tests. But the network score obtained is much less than that of Greedy algorithm as hill climb algorithm is not an exhaustive algorithm [2].

22 e) Greedy Algorithm

A greedy algorithm is an algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding a global optimum. The network score obtained is much higher as greedy algorithm is an exhaustive algorithm. Because of this we got the desired Bayesian Network [2]. [1] [2] [3]

2013

Figure 1: O © 2013

Figure 2:

[1] © 2013 Global Journals Inc. (US)

[2] DBayesian Network Model for Epidemiological Data

[3] © 2013 Global Journals Inc. (US) Global Journal of Computer Science and Technology

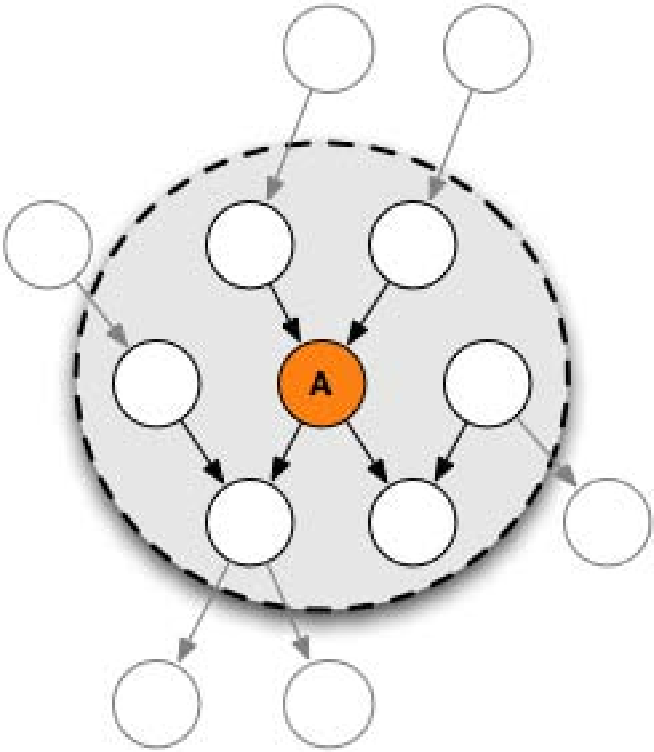

Figure 3: Year

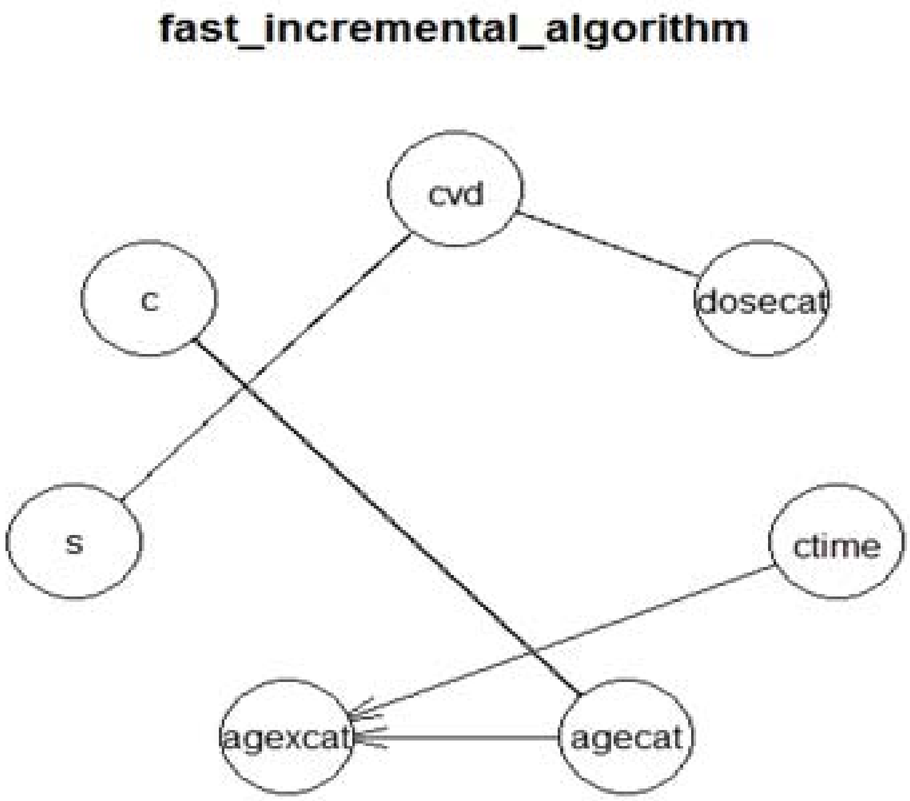

Figure 4: D

So from the above discussion we can understand that the Greedy algorithm is the best algorithm to implement the Bayesian Network for the given dataset which considers all the possible relationship between the variables and finds a complete Bayesian network(Note: It is not mandatory that the same algorithm is best suitable for all the data.). Also, using the probability distribution obtained from above network, we found that radiation dosage below 0.5 Gy (dosecat 0-13) have negligible effect on death count (CVD, STROKE, HEARTD). The above three are the probabilistic distribution models for CVD, STROKE and Heart D.

[Heckerman ()] *A tutorial on learning with Bayesian networks*, D E Heckerman . MSR-TR-95-06. 1996. Redmond, WA, Microsoft Research.

[Tu ()] 'Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes'. J V Tu . *JClinEpidemiol* 1996. 49 p. .

[Heckerman] 'Bayesian networks for data mining'. D Heckerman . *Data Min Knowledge Discovery997* 1 p. .

[Hamilton et al. ()] 'Clinical applications of Bayesian belief networks in pathology'. P W Hamilton , R Montironi , W Abmayr . *Pathologica* 1995. 87 (3) p. .

[Luis Targo (2011)] *Data mining with R*, Luis Targo . `http://www.crcp-ress.com` April 2011. Chapman and Hall, CRC press.

[Bøttcher (2003)] *deal: A Package for Learning Bayesian Networks*, Susanne G Bøttcher , Clausdethlefsen . April 2003. (bayesian network with deal)

[Susanne et al. (2003)] 'Learning Bayesian Networks with R'. " Susanne , G Bøttcher , Claus Dethlefsen . `http://www.ci.tuwien.ac.at/Conferences/DSC-2003` *3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 2003.

[Scutari (2009)] 'Learning Bayesian Networks with the bnlearn R Package'. Marco Scutari . *Journal of statistical software* June 2009. (bayesian network with bnlearn)

[Shimizu ()] 'Life Span Study Circulatory Disease Mortality Data'. Y Shimizu . `http://www.rerf.or.jp` *RERF foundation*, 1950-2003. 2010.

[Cios ()] *Medical data mining and knowledge discovery*, K J Cios . 2001. New York: Physica-Verlag.

[Li et al.] 'Modeling and Analysis of Disease and Risk Factors through learning Bayesian Network from Observational Data'. Jing Li , Jianjun Shi , Devin Satz ; Susanne , G Bottcher . *Marco Scutari* 12. (Deal: A Package for Learning Bayesian Networks. Susanne G. Bottcher, Claus Death lefsen)

[Li et al. (2007)] 'Modeling and Analysis of Disease and Risk Factors through Learning Bayesian Networks from Observational Data'. "jing Li , Jianjun Shi , Devin Satz . `www.interscience.wiley.com` *WileyInter Science* November 2007.

[Shimizu ()] *Radiation exposure and circulatory disease risk: Hiroshima and Nagasaki atomic bomb survivor data*, Yukiko Shimizu . 2010. p. . (Research paper on survivor data)