Bayesian Methods for Diagnosing Physiological Conditions of Human Subjects from Multivariate Time Series Biosensor Data

Mehmet Kayaalp

Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland
Email: Mehmet.Kayaalp@NIH.gov

In this study, we developed a set of new methods to diagnose physiological conditions (PCs) of human subjects based on biometric multichannel time series data that were obtained through nine biosensors attached on the upper arms of the human test subjects. The task is to learn the relationships between PCs and the biosensor data on the training dataset and identify the PCs of interest on the test dataset.

The goal of this study is to test the predictive performances of two sets of Bayesian network topologies and the effectiveness of a new parameter learning method applied to the problem on the provided dataset. This paper summarizes these methods.

Data

The dataset was collected and made available to the public by BodyMedia, Inc. as the workshop dataset of the Physiological Data Modeling Contest at the International Conference on Machine Learning in 2004.[1] The entire dataset was deidentified such that the identifiers of human subjects and the original labels of PCs were replaced with unique integers.

The training dataset was provided in 16 columns (see Table 1) and 580,264 cases (temporal records). The test dataset comprises 720,792 cases, whose format follows that of the training dataset except for the columns *PC* and *gender*, which are omitted and left to be predicted by the participants for each test case.

Table 1: The Format of the Training Dataset[2]

Line	***char*1**	***char*2**	***PC***	***gender***	***sensor*1**	⋯	***sensor*9**	***userID***	***sessionID***	***sessionTime***
1	38	0	0	0	0.168662	⋯	-0.41966	1	2	0
2	38	0	0	0	0.159785	⋯	-0.31002	1	2	60000
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
580264	24	0	0	0	0.754541	⋯	-0.30018	32	4064	5400000

[1] The dataset decomposed into two mutually exclusive subsets, training and test data, is available at the official website of the workshop: http://www.cs.utexas.edu/users/sherstov/pdmc/

[2] For the semantics of *char*1, *char*2 and *gender*, please refer to Section **Determining Genders of the Subjects**.

Dynamic Simple Bayesian Modeling

Dynamic simple Bayesian (DSB) modeling is a Bayesian method to forecast future outcomes of stochastic processes given a sequence of multivariate time series (Kayaalp, 2003). A DSB model is a dynamic Bayesian network with the following assumptions (see Figure 1):

1. The underlying data-generating stochastic process is strictly stationary, which implies that the parameters of the stochastic process are constant under any time displacement $d \in \mathbb{Z}$; i.e., $P\left(\mathbf{X}(t_i),...,\mathbf{X}(t_n)\right) = P\left(\mathbf{X}(t_{i+d}),...,\mathbf{X}(t_{n+d})\right)$.
2. All contemporaneous variables of a stochastic process are measured concurrently, and all successive measurements are performed in equidistant time intervals.
3. Temporal dependencies are first-order Markov; that is, the state of a process at time t_n depends only on the state of the same process at time t_{n-1}, i.e. $P\left(\mathbf{X}(t_n)\middle|\mathbf{X}(t_{n-1})\right) = P\left(\mathbf{X}(t_n)\middle|\mathbf{X}(t_{n-1}),...,\mathbf{X}(t_0)\right)$.
4. The observations on a given time slice *t* are conditionally independent, given the outcomes in $t+1$.
5. All variables are discrete.

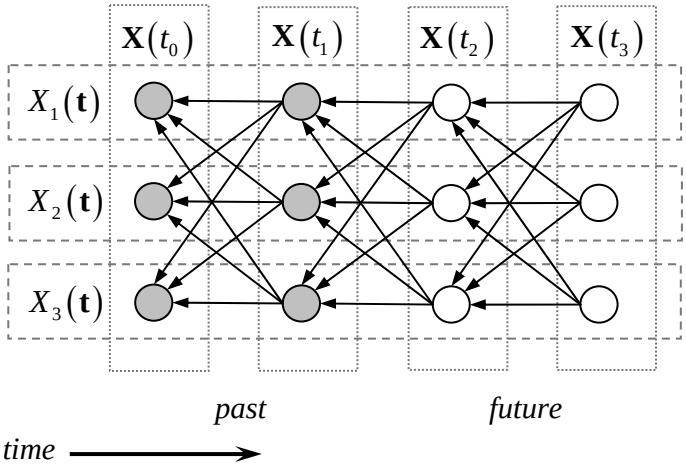

Figure 1: Dynamic Simple Bayesian (DSB) Model with Three Temporal Variables

In the provided dataset, physiological signals are represented in continuous numbers over a discrete timeline. The conventional approach of using continuous data in Bayesian networks usually requires either a priori discretization of the data or a priori assumption that the data conform to a particular parametric distribution such as Gaussian. These approaches may be optimal under certain conditions; however, they may fail when their (rather strong) premises contradict with the underlying distributions.

In this study, parameters of the Bayesian networks are learned from continuous data. Consider the Bayesian network in Figure 2, where *X* and *Y* are continuous and binomial variables, respectively.

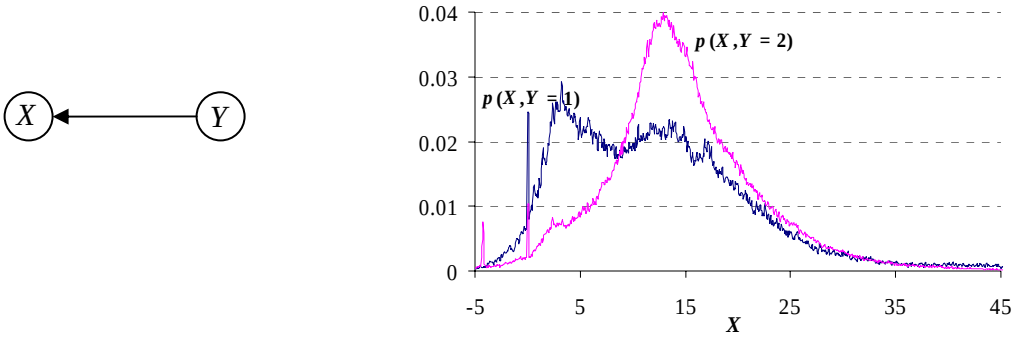

Figure 2: A Bivariate Bayesian Network and the Associated Density Distributions

The area under any segment of a probability density function renders the probability density of that segment,

$$p(x_1 \le X \le x_2) = \int_{x_1}^{x_2} p(X)\,dx, \tag{1}$$

and the probability of *X* at any given point x is equal to 0. [3]

$$P(X = x) = \int_{x}^{x} p(X)\,dx = 0 \tag{2}$$

Using continuous data in Bayesian networks has always been problematic. The most important problem is the difficulty of obtaining the integral value of Equation (1) for multivariate distributions. The probability of any point on the real line is a null set may also be construed as a hurdle to parameterize a test data point in question. To solve the current problem, what we need is to obtain the value of $P(Y|X = x)$, which is a non-zero value for a continuous variable *X* and a categorical variable *Y*. Namely, in the example illustrated in Figure 2, $P(Y = 1|X = x) + P(Y = 2|X = x) = 1$.

Given a test case in which $X = x$, how can we learn $P(Y = y|X = x)$ from a finite sample? Our approach is online parameter learning; that is, parameters of every test case are learned dynamically. Each probability of interest is learned parametrically by combining values of all training sample points via a sigmoid decay function (sdf), which discounts the contribution of each sample point x_i according to the Euclidean distance between x_i and the test data point x of interest.

[3] In this work, we denote probability density functions with (a lower case letter) *p* and probability mass functions with (an upper case letter) *P*.

$$N_{y_k|x} = \sum_i \frac{n(y_k, x_i)}{1 + e^{\frac{a}{scale(X)}|x_i - x| - c}} \tag{3}$$

$$P(y_k|x) = \frac{\alpha_k + N_{y_k|x}}{\alpha + \sum_{y \in Y} N_{y|x}} \tag{4}$$

In Equation (3), $n(y_k, x_i)$ denotes the number of data points in the training sample where values of variable *Y* and *X* are equal to y_k and x_i, respectively, and $scale(X) = \max(X) - \min(X)$. In this study, we set sdf parameters as $a = 400$ and $c = 6$. In Equation (4), α_k / α denotes the prior probability of $Y = y_k$ given $X = x$, and $\alpha = \sum_i \alpha_i$, where all priors are distributed uniformly, i.e. $\forall (i, k | i \neq k)\ \alpha_i = \alpha_k$.

The computational complexity of Equation (3) as implemented in this study is $\Theta(n^2)$ where *n* is the number of unique[4] data points in the training sample. However, the current algorithm may easily be approximated numerically to achieve $\Theta(1)$.

This method enables us to eliminate discrete data requirement (Assumption 5) from the list of assumptions without making as strong assumptions about the underlying distributions of the data as done by other methods (Shachter & Kenley, 1989; Cowell, 1998). The rationale of the method is based on the natural order of the continuous data. It takes into account that measurements of a given phenomenon are usually bound to a certain variance, which may be accommodated through the parameters of the sdf.

In this study, the outcomes of interest are categorical; hence, we restrict our implementation to discrete *Y* variables. If *Y* were defined on real line, the two density plots in Figure 2 would be a surface plot. In that case, the unidimensional Euclidian distance metric $|x_i - x|$ of Equation (3) would be replaced with a two-dimensional one such as $\sqrt{(x_i - x)^2 + (y_j - y)^2}$ and the single summation over *i* in Equation (3) would be replaced with a double summation over *i* and *j*. Extending sdf to an *n*-dimensional metric space is straightforward (Berge, 1962).

Modified Dynamic Simple Bayesian Modeling

DSB models are designed for forecasting (i.e. for predicting future outcomes); however, the current task is diagnosing physiological conditions at time *t* from time series data that include sensor information at time *t*. Such contemporaneous information is too valuable to ignore;

[4] Depending on sensor resolutions, actual data do not always reflect the analog nature of the measured signals; rather, certain data points appear quite frequently. For instance, each of the following Sensor 3 data points 32.582733 and 32.751022 was recorded more than 1000 times in the training dataset. Resolutions of Sensors 6 and 8 seem significantly higher than the others; thus, computing Equation (3) took much longer for the data of these two sensors than for the data of the others.

thus, we modified current DSB model and added contemporaneous variables. The modified DSB model (mDSB) can be conceptualized as a coalescence of a DSB model and a simple Bayesian network model (SB); the latter is also known as naive Bayes classifier.

An SB model is essentially an atemporal model such that its random variables are not associated with any temporal semantic and variable interdependencies do not indicate any temporal transitions. In this study, the following assumptions are made for using the SB approach in the temporal domain:

1. An event at time t_n are independent from any other event at time $t \neq t_n$; that is,

$$P\left(\mathbf{X}(t_n)\middle|\mathbf{X}(t_{n-1}),...,\mathbf{X}(t_0)\right) = P\left(\mathbf{X}(t_n)\right). \quad (5)$$

2. All covariates are conditionally independent given the outcome of the variable of interest.

The main difference between the DSB and SB models is that the former contains only transitional (temporal) relations whereas the latter contains only contemporaneous ("atemporal") relations. As a combination of the two, an mDSB contains both transitional and contemporaneous relations.

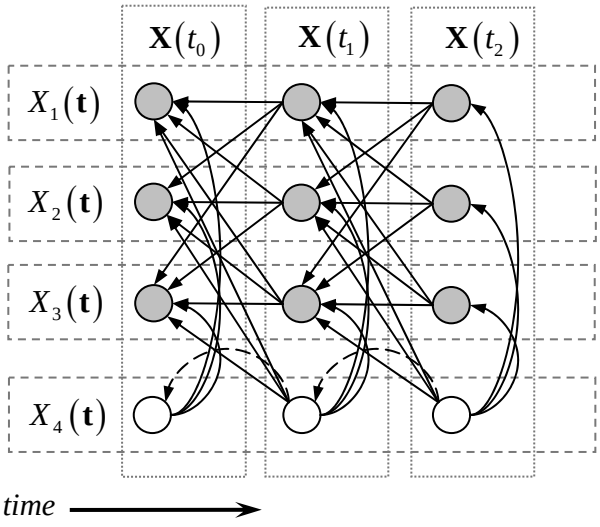

Figure 3: An mDSB Model

The SB portion of an mDSB model comprises a set of variables $\mathbf{X}(t_i)$ and a set of relations contemporaneous to t_i (see Figure 3). The values of the variables of interest denoted by unshaded nodes were unobserved (i.e., unknown). In the current study, transitional relations (represented with the dashed arcs in Figure 3) between successive unknown variables were not implemented. In the general case of mDSB, variables with unknown values may also

have a complete set of first-order Markov dependencies as their observable contemporaneous counterparts.

Determining Genders of the Subjects

There are three additional variables of the training dataset that are not mentioned above: The gender of each subject, and two sets of subject characteristics, which we here denote as *char*1 and *char*2 (see Table 1). All three variables were deidentified and enumerated. Here are some observations about this portion of the training dataset:

- Both gender and *char*2 were static.
- Changes in *char*1 over time were rare.
- Effective sample size was small—four subjects with *gender* = 1 and 14 subjects with *gender* = 0 in the training dataset.

The task specified by the organizers of the workshop was to determine the gender of each test subject by using all available information, which includes time series sensor data as well. With only a small set of parameters, we identified a deterministic relation between gender and the two characteristics on the training dataset through constructive induction:

$$gender = \begin{cases} 1 & \text{if } \ char2 = 1 \vee char1 = 29 \vee char1 = 42 \\ 0 & \text{otherwise} \end{cases}$$

Results and Conclusions

In this study, we applied a set of new Bayesian methods to diagnose physiological conditions (PCs) of human subjects from multivariate time series data that were obtained through a set of biosensors. We submitted our results in three sets: The first two sets were predicted by two mDSBs with two different decision thresholds and the last set was predicted by an SB. At the time of submission of this paper, actual genders and PC values of the subjects in the test data were not publicized; thus, we defer the evaluation of the methods until the availability of the test results.

References

Berge, C. (1962). *Espaces topologiques, fonctions multivoques* (*Topological Spaces* translated by E.M. Patterson. Dover, Mineola, NY, 1997 ed.). Paris, France: Dunod.

Cowell, R. (1998). Advanced Inference in Bayesian Networks. *Learning in Graphical Models* (p. 27–49). Boston, MA: Kluwer Academic Publishers.

Kayaalp, M. M. (2003). *Learning Dynamic Bayesian Network Structures from Data.* Ph.D. Dissertation, University of Pittsburgh, Pittsburgh, PA.

Shachter, R. D., & Kenley, C. R. (1989). Gaussian Influence Diagrams. *Management Science, 35*(5), 527–550.