Group 8: Multivariate time-series prediction of ICU mortality

Paper Review

Danning He

April 2nd, 2012

Paper Selection:

Saria S, Rajani AK, et al. (2010). "**Integration of early physiological responses predicts later illness severity in preterm infants.**" Sci Transl Med. 2(48): 48ra65.

Overview:

The paper developed a new scoring system "PhysiScore" using 8 physiological variables collected during first 3 hour of the NICU to classify 138 preterm neonates into to high morbidity (HM) group and low morbidity (LM) group. The innovation in this study is that it transforms multiple observations associated with each physiological variable into a numeric risk feature via nonlinear Bayesian modeling, and then aggregated individual risks features using standard logistic regression classifier. Under this probabilistic framework, the method shows superior classification performance compared with several popular, state-of-art scoring systems.

Nonlinear models of risk features

Use parametric model to represent probabilities

This ratio will be 1 if variable is observed

$$f(v_i)=\begin{cases}\log\dfrac{P(v_i \mid HM, m_i=0)}{P(v_i \mid LM, m_i=0)}\cdot\dfrac{P(m_i=0 \mid HM)}{P(m_i=0 \mid LM)} \\ \log\dfrac{P(m_i=1 \mid HM)}{P(m_i=1 \mid LM)}\end{cases}$$

The likelihood ratio that the measurement is missing in HM versus LM

The nonlinear model of risk features is the most innovative part of this paper, so I'll mainly focus on discussing this approach. In this study, various physiological variables are observed at multiple time points, but the

predictive factor v_i is defined as the derivative statistics of the original observations (such as the mean, variation, percent of time below a certain threshold). And for each predictive factor, it learned a parametric model of the distribution of all observed values for each class of patient in the training set. To allow more flexibility in curve fitting, the parametric model is selected and learned with maximum-likelihood estimation from five long-tailed probability distributions (exponential, Weibull, lognormal, normal, and gamma). Although exponential distribution has only one parameter, while the other four have two parameters, the use here only allows better fitting. Another point to notice is that, such approach takes into account the possibility that the overall behavior of a predictive factor can behave differently between sickness categories by allowing the same factor to fit with different distribution in two classes (e.g., a predictive factor can belong to exponential distribution in high morbidity class belong to gamma distribution in low morbidity class). And the log odds ratio of the predictive factor was incorporated in to the model.

A very important advantage of this approach is that it allows explicit assumptions of missing observations. Unlike the measurements of some standard physiological indices such as heart rate, blood pressure, some laboratory tests (such as blood gas measurements) are performed only when the patients are deemed to be at higher risk, and hence are not missing at random, therefore the likelihood that a particular measurement might be taken for a patients in different categories is also incorporated into the model. The presence ($m_i = 0$) or absence ($m_i = 1$) of the measurement itself is also informative under this condition.

Yet another advantage of the parametric representation is that it alleviates issues arising from data scarcity. It is assumed that the measurements are samplings from an underlying continuous-valued state space, and it's impossible to observe all possible values. Such approach can assign non-zero probabilities to unobserved values that may appear in the future.

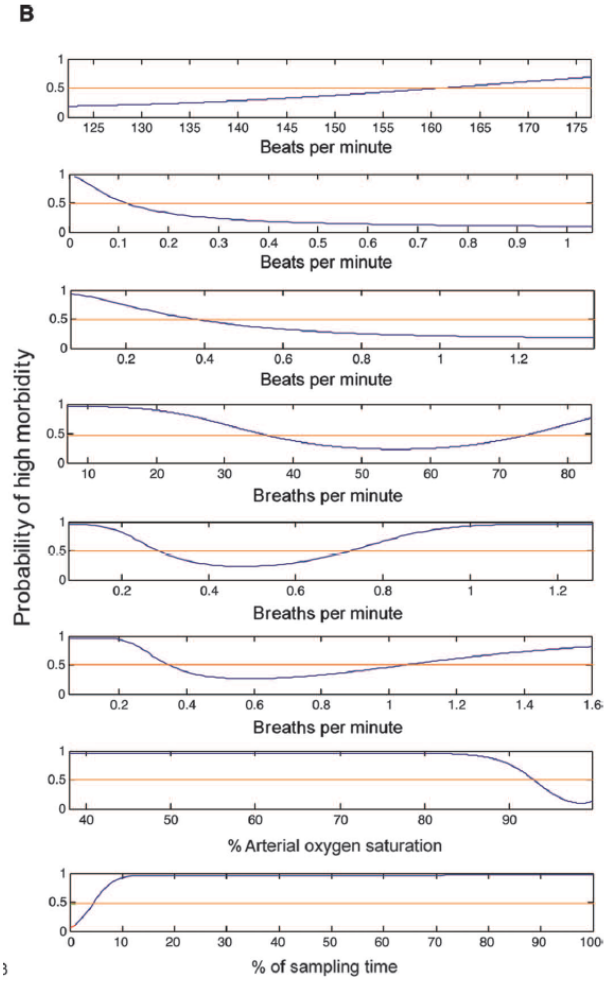
Aggregate individual risk features using logistic function

The logistic function is widely used in statistics and medical research to model how the probability P of an even (with range in [0,1]) may be affected by one or the linear combinations of multiple explanatory variables, which may be of any type: real-valued, binary and categorical, etc.

$$P(HM \mid v_1, v_2, ..., v_n) = \left(1 + \exp\left(b + w_0 \cdot c + \sum_{i=1}^{n} w_i \cdot f(v_i)\right)\right)$$

where n was the number of predictive factors (a physiological variable can have several forms of derivative statistics such as mean, variations and etc.), and $c = \log \frac{P(HM)}{P(LM)}$ was the prior log odds ratio.

The score parameters b and w were learned from the training data via maximum likelihood, The parameter $\mathbf{w_i}$ represents the weight of the contribution of the $\mathbf{i}_{th}$ characteristic to the computed probability score, with higher weight characteristics having a greater effect.



Results:

Besides several routines for classification evaluation such as receiver operating characteristic (ROC) curve and associated area under the curve (AUC) values, the paper also obtained an interpretable visualization of the likelihood of low patient morbidity over the range of values for each features. In the figure above, horizontal-axis represents the range of values, and the vertical axis represents the probability of high morbidity (HM) conditional on a particular value. For example, for the fourth subfigures from the top "breaths per minute", P(HM)<0.5 only within the range from 35~75. When the value is less than 35 or larger than 75, the P(HM | value) goes above 0.5. But notice that P(HM | value) doesn't necessarily span the whole [0,1] interval.

Critique:

Although this paper is very similar to my project, there's a very important difference that is not obvious at first glance. The paper classifies patients into high and low morbidity, instead of mortality. Here in this paper, morbidity was defined by moderate or severe bronchopulmonary dysplasia (BPD),

retinopathy of prematurity (ROP) stage 2 or greater, intraventricular hemorrhage (IVH) grade 3 or 4, and necrotizing enterocolitis (NEC), each of which is dysfunction in a certain organ. And of the 37 out of 138 patients with high morbidity, only 4 of them died. It means the result won't be so perfect if the goal is to predict mortality—it will have higher false positive rate. If the causal relationship is from abnormal physiological profiles->high morbidity->high mortality, this paper deals with a simpler task than our project. The real complexity in our project is that, mortality is influenced not only by inner status of the patients, but also by external interference—such as a successful/unsuccessful surgery, which is unknown in our data.

Another difference is the sample size, the study consists of 138 patients in total, therefore, even only 8 physiological variables are predictive enough. Small sample size also enables the use of the powerful leave-one-out cross validation (LOOCV), which becomes computationally prohibitive in large dataset, as is the case in our project (4000 samples in total).

One possible improvement to this paper is to incorporate the counts of laboratory tests in defining the risk feature as well. In the original method, it only considers whether a certain physiological variable is missing or not ($m_i = 1$ or $m_i = 0$), and doesn't take into consideration how many times it is observed. But a point to notice is that incorporating the counts will also arise the issue of whether there're dependencies between the counts of tests and the values of tests. Another improvement to the method is to explicitly incorporate the values of multiple observations. In the original paper, it transforms the vector-valued observations within first 3 hour into a scalar-valued derivative statistics (such as mean, variance). Our dataset has more observations per variables which lasts for longer time (48 hour), using the likelihood of all observations better retains the information, although in both study, the timestamp of observations is not incorporated into the model.