Early Prediction of Sepsis from Clinical Data Using a Specialized Hidden Markov Model

Supplementary Abstract

An HMM is a state-space model which consists of a hidden process \( \{C_t; t = 1, 2, \ldots\} \), called states, which is a Markov chain of order 1, and an state dependent observed process \( \{X_t; t = 1, 2, \ldots\} \), called observations, with the joint distribution of observations and states in the time period \( t = 1, \ldots, T \) as

\[
P(X_{1:T}, C_{1:T}) = P(C_1) \prod_{t=2}^{T} P(C_t|C_{t-1}) \prod_{t=1}^{T} P(X_t|C_t).
\]

For the sepsis data set, the first 34 columns, \( Y_t \), can be considered as continuously distributed observations, while the "SepsisLabel", \( W_t \), is a binary variable. The observation model would be

\[
P(X_t|C_t) = P(Y_t|C_t, W_t) \times P(W_t|C_t).
\]

A 3-state model is considered with three hidden states "healthy=1", "ill=2" and "sepsis=3". Thus, letting \( Q_{ij} = P(W_t = j - 1|C_t = i), \ i = 1, 2, 3, \ j = 1, 2 \), it is trivial that the matrix \( Q = ((Q_{ij})) \) is given by

\[
Q' = \begin{pmatrix}
1 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]

A Gaussian model is considered for \( P(Y_t|C_t, W_t) \) as

\[
P(Y_t|C_t = i) = \mathcal{N}(Y_t; \mu_i, \Sigma_i),
\]

with mean vectors \( \mu_i \) and diagonal variance-covariance matrices \( \Sigma_i, \ i = 1, 2, 3 \). An expectation-maximization (EM) algorithm is used to estimate the parameters, while the prior probabilities are assumed to be equal to \( (1, 0, 0)' \) and the restrictions \( \Gamma_{1,3} = \Gamma_{3,1} = \Gamma_{3,2} = 0 \) and \( \Gamma_{3,3} = 1 \) are imposed to the transition matrix \( \Gamma \).

The missing observations are treated as non-observed variables in the proposed EM algorithm. Based on a sample, the final estimates of each parameters is obtained as the mean of the final converged value. To obtain the initial values, a specialized imputation method is first applied to each subject. Then, the imputed observations are clustered using some specialized ordered k-means algorithm. To predict the state in a given time \( t + h, h \geq 1 \), the conditional probabilities \( P(C_{t+h} = j|X_{1:t}) \) are computed. A 3-fold cross-validation on a subset of the train data set resulted in a maximum cross-validated utility is almost equal to 0.88.

\[1\]

Morteza Amini, Email: morteza.amini@ut.ac.ir, Department of Statistics, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran.