A new approach in classification discrete survival data by discrete hidden Markov model and comparison to one common method

Farzan MADADIZADEH *PhD student in Biostatistics, School of Public Health, Tehran University of Medical Sciences, Iran*, E-mail: fmadadizadeh@razi.tums.ac.ir
Hojjat Zeraati *Professor of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Iran*
Vahid rezaietabar *Department of Mathematics and Statistics, Faculty of Financial Science, University of Economic Sciences, Iran*
Abbas Bahrampour *professor of Biostatistics, Department of Epidemiology and Biostatistics, Kerman University of Medical Sciences, Iran*

Background: Diagnosis of cancer survival is very important and existence of reliable system is necessary for reduce the medical expenditure. Discrete Hidden Markov Models (DHMMs) are a ubiquitous tools and probabilistic techniques for modeling sequence data. These models include an underlying stochastic process that is hidden but could be inferred through the observations it generates. Discrete time Survival data comprised of a set of events and observation during the discrete time. This paper presents the performance of DHMM in classification of event and compere results with logistic regression that considers overall probability of occurrence of event.
Methods: A dataset was acquired of Health Department and Cancer registry of Kerman Province which is located in south of Iran, and data includes the information of 900 breast cancer patients aged 15 to 80 years, and its seven associated risk factors among patients since 1999 to 2007. Our breast cancer data are assumed to include two possible events such as live and died. DHMM was trained based on a set of 675 patients information (75% of data) and that was validate in a test set of 225 patients information (25% of data). The Area Under the ROC Curve (AUC), sensitivity, specificity and accuracy used as measures of validate of the efficiency of model in prediction of patients survival status.
Results: Sensitivity, specificity, accuracy and the and area under the ROC curve of the logistic regression was 0.86, 0.97,0.92 and 0.951, respectively. The sensitivity, specificity, accuracy and the area under the ROC curve of DHMM was 0.989, 0.99, 0.939 and 0.94, respectively.
Conclusions: According to the four evaluation criteria, DHMM would give better performance than logistic regression.

Acknowledgement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Fink GA (2014). Hidden Markov Models. Markov Models for Pattern Recognition, *Springer*, 71-106.

[2] Rabiner L (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77**(2), 257-86.

[3] Hassan MR, Hossain MM, Begg RK, Ramamohanarao K, Morsi Y (2010). Breast-cancer identification using HMM-fuzzy approach, *Computers in biology and medicine*, **40**(3), 240-51.