DSSV 2023 DATA SCIENCE, STATISTICS & VISUALISATION
ECDA 2023 EUROPEAN CONFERENCE ON DATA ANALYSIS

Modeling Multiple-Criterion Diagnoses by Heterogeneous-Instance Logistic Regression

C.-H. Yang[1], M.-H. Li[2], S.-F. Wen[2], and S.-M. Chang[3]

[1]National Taiwan University, Institute of Applied Mathematical Sciences, Taipei, Taiwan
[2]Industrial Technology Research Institute, Biomedical Technology and Device Research Laboratories, Hsinchu, Taiwan
[3]National Taipei University, Department of Statistics, New Taipei, Taiwan

The diagnoses of Alzheimer's disease (AD) and its prodromal stage mild cognitive impairment (MCI) are examples of multiple-criterion diagnoses. Clinically, the diagnosis of MCI/AD is determined by the impairment statuses of five cognitive domains. If one of these cognitive domains is claimed impaired, the patient is diagnosed with MCI, and if two out of the five domains are impaired, the patient is diagnosed with AD. This diagnostic procedure relates the MCI/AD status modeling to multiple-instance learning [1], where each domain resembles an instance. However, traditional multiple-instance learning assumes common predictors among instances, but in our case, each domain is associated with different neuropsychological questionnaires [2] as predictors. In this paper, we generalized the multiple-instance logistic regression to accommodate the heterogeneity in predictors among different instances. The proposed model is dubbed Heterogeneous-Instance Logistic Regression (HILoR) and the expectation-maximization algorithm[3] is applied to solve the estimation problem arising from the missing instance statuses. We also derived two variants of HILoR to model both the MCI diagnosis and AD diagnosis. Because HILoR fully considers the probability structure of a multiple-criteria diagnosis, the proposed model additionally provides the probabilities of being impaired in each criterion/domain. Take our data analysis for example. We predicted one's MCI (AD) status after 3 (5) years of the initial visit and then provided the impairment probabilities (or percentiles) in each of the five domains. Treatments can therefore be applied according to more deteriorated domains. This merit contributes to personalized disease prevention. Finally, the proposed HILoR model is validated in terms of its estimation accuracy, latent status prediction, and robustness via extensive simulation studies and real data analysis.

Keywords: EM Algorithm, Logistic Regression, Multiple-Criteria Diagnosis, Multiple-Instance Learning

[1] R. Chen, K. Cheng, S. Chang, S. Jeng, C. PY, C. Yang, and C. Hsia, "Multiple-instance logistic regression with lasso penalty," *arXiv*, p. 1607.03615, 2016.

[2] M. Albert, S. DeKosky, D. Dickson, B. Dubois, H. Feldman, N. Fox, A. Gamst, D. Holtzman, W. Jagust, R. Peterson, P. Snyder, M. Carrillo, B. Thies, and C. Phelps, "The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & Dementia Journal*, vol. 7, no. 3, pp. 270–279, 2011.

[3] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.