Accepted Manuscript

Classified mixed logistic model prediction

Hanmei Sun, Thuan Nguyen, Yihui Luan, Jiming Jiang

PII: S0047-259X(18)30076-9
DOI: https://doi.org/10.1016/j.jmva.2018.06.004
Reference: YJMVA 4372

To appear in: *Journal of Multivariate Analysis*

Received date : 4 February 2018

Please cite this article as: H. Sun, T. Nguyen, Y. Luan, J. Jiang, Classified mixed logistic model prediction, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.06.004

Classified mixed logistic model prediction

Hanmei Sun[a], Thuan Nguyen[b], Yihui Luan[a], Jiming Jiang[c]

[a]*School of Mathematics, Shandong University, Jinan, Shandong 250100, China*
[b]*Oregon Health and Science University, Portland, OR 97239-3098, USA*
[c]*Department of Statistics, University of California, Davis, CA 95616, USA*

Abstract

We develop a classified mixed logistic model prediction (CMLMP) method for clustered binary data by extending a method proposed by Jiang et al. (2018, J. Amer. Statist. Assoc.) for continuous outcome data. By identifying a class, or cluster, that the new observations belong to, we are able to improve the prediction accuracy of a probabilistic mixed effect associated with a future observation over the traditional method of logistic regression and mixed model prediction without matching the class. Furthermore, we develop a new strategy for identifying the class for the new observations by utilizing covariates information, which improves accuracy of the class identification. In addition, we develop a method of obtaining second-order unbiased estimators of the mean squared prediction errors (MSPEs) for CMLMP, which are used to provide measures of uncertainty. We prove consistency of CMLMP, and demonstrate finite-sample performance of CMLMP via simulation studies. Our results show that the proposed CMLMP method outperforms the traditional methods in terms of predictive performance. An application to medical data is discussed.

Keywords: Clustered binary data, CMLMP, CMMP, Matching, Mixed logistic model, Mixed model prediction, MSPE

1. Introduction

Many practical problems are related to prediction of characteristics of interest at subject (e.g., precision medicine) or sub-population (e.g., precision public health) level. In such cases, it is possible to make substantial gains in prediction accuracy by identifying a class that a new subject belongs to. This was recently demonstrated by Jiang et al. [7], who proposed a method called classified mixed model prediction (CMMP). The idea is to create a "match" between the classes, or clusters, of the training data and the potential class of the new data. Once such a class is identified, a mixed model prediction (MMP) technique can be used to improve prediction accuracy; see, e.g., Section 2.3 in [4].

The CMMP method developed in [7] applies only to linear models for continuous responses. However, clustered binary data frequently occur in practice. For example, Thromboembolic or hemorrhagic complications [2] occur in as many as 60% of patients who underwent extracorporeal membrane oxygenation (ECMO), an invasive technology used to support children during periods of reversible heart or lung failure [8]. Over half of pediatric patients on ECMO are currently receiving antithrombin (AT) to maximize heparin sensitivity. In a retrospective, multi-center, cohort study of children ($\leq$ 18 years of age) who underwent ECMO between 2003 and 2012, 8601 subjects participated in 42 free-standing children's hospitals across 27 US states and the District of Columbia known as Pediatric Health Information System (PHIS). Data were de-identified prior to inclusion in the study dataset; however, encrypted medical record numbers allowed for tracking of individuals across multiple hospitalizations. Many of the outcome variables were binary, such as the bleed_binary variable, which is a main outcome variable indicating hemorrhage complication of the treatment; and the DischargeMortalit1Flag variable, which is associated with mortality. Here the treatment refers to AT. Prediction of characteristics of interest associated with the binary outcomes, such as probabilities of hemorrhage complication or those of mortality for specific patients are of considerable interest. Note that the data are also potentially clustered, with the clusters corresponding to the children's hospitals. In addition to the treatment indicator, there were 20 other covariate variables, for which information were available. More details about the data will be provided in Section 5.