Computational Statistics and Data Analysis 94 (2016) 238–249

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A sequential logistic regression classifier based on mixed effects with applications to longitudinal data

Xin Zhang[a], Daniel R. Jeske[a,*], Jun Li[a], Vance Wong[b]

[a] Department of Statistics, University of California—Riverside, Riverside, CA 92521, USA
[b] Alere Inc., San Diego, CA 92121, USA

ARTICLE INFO

Article history:
Received 16 May 2014
Received in revised form 18 August 2015
Accepted 18 August 2015
Available online 28 August 2015

Keywords:
Classification
Mixed effects models
Longitudinal data

ABSTRACT

Making an early classification in longitudinal data is highly desirable. For this purpose, a sequential classifier that incorporates a neutral zone framework is proposed. The classification procedure evaluates each subject sequentially at each longitudinal time point. If there is not adequate confidence in making a classification at a given time point, the decision will wait until the next time point where another measurement is collected. This process continues until there is enough confidence of making a classification or until the last time point where data can be collected is reached. It is demonstrated that the proposed sequential classifier maintains competitive error rates while reducing the overall cost when the cost of time is taken into account. The classifier is applied to a real example of identifying patients that are vulnerable to kidney dysfunction on the basis of up to 7 blood draws sequentially taken from each patient.

1. Introduction

Statistical classification is the problem of assigning an observation to one of a set of populations, based on a training set of observations whose population membership is known. One applicable area of statistical classification is disease diagnosis. Many diseases can be detected based on the difference in levels of certain clinical characteristics (e.g. biomarkers) between the disease group and the non-disease group. Commonly, these characteristics would be repeatedly measured throughout hospitalization, so that disease diagnosis can be made according to patients' profiles. Consider, for example, the data illustrated in Fig. 1, which consists of biomarker profiles in a kidney dysfunction study. The data includes 319 patients, 40 of which developed kidney dysfunction within 72 h following heart surgery and 279 who did not. Within a 72 h window, as many as 7 blood draws were taken from each patient at time points $t = 0, 3, 6, 12, 24, 48$, and 72 h, and the biomarker level at each time point was recorded. Here, time 0 corresponds to the point at which the patient had the surgery. The group labels, "dysfunction" and "no dysfunction", were determined based on the condition that ultimately developed within 72 h for each patient. For this application, the goal is to use these data as a training data set to build a classifier that determines if the evolving trajectory of a new heart surgery patient is tracking with patients that ultimately developed kidney dysfunction.

Classifying longitudinal data presents more challenges than classifying regular multivariate data. First, there are usually many missing values in the longitudinal data. Second, the time points at which the repeated measures are observed may vary between subjects. To overcome these difficulties several procedures have been proposed for classifying longitudinal data (see Verbeke and Lesaffre, 1996, Marshall and Barón, 2000, James and Hastie, 2001, Luts et al., 2012 for example).

[*] Correspondence to: 1340 Olmsted Hall, Department of Statistics, University of California—Riverside, Riverside, CA, 92521, USA.
E-mail address: daniel.jeske@ucr.edu (D.R. Jeske).

http://dx.doi.org/10.1016/j.csda.2015.08.009
0167-9473/© 2015 Elsevier B.V. All rights reserved.

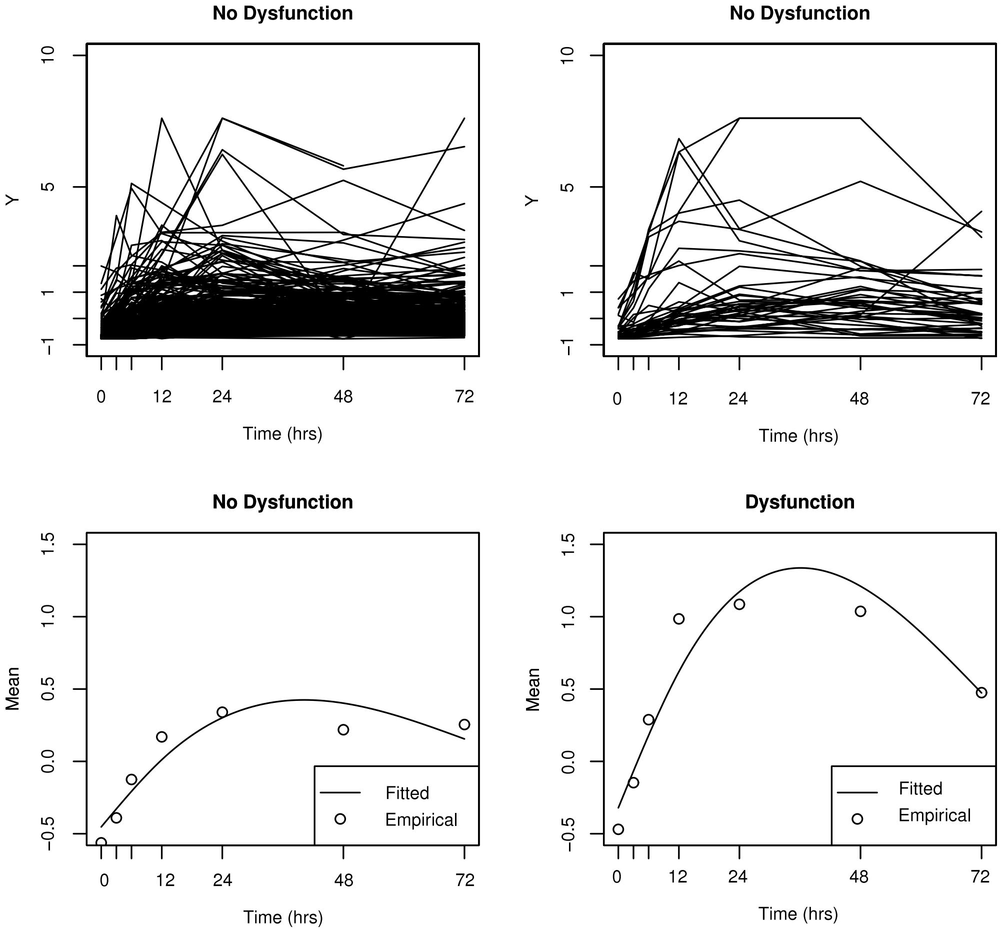

Fig. 1. Biomarker measurements for kidney data.

These procedures were developed to classify the longitudinal data based on the complete profile of the data. In the kidney dysfunction study, those procedures would classify the subject at the end of 72 h. However, it is extremely desirable to classify the subject as early as possible, since kidney dysfunction is a life-threatening condition and an early indicator of its impending likelihood would result in a lower mortality rate. If existing longitudinal data classifiers were simply applied at earlier time points, the performance of the classifiers may be compromised, since some subjects may not have enough information to be correctly classified at earlier time points. On the other hand, some subjects may indeed show signs of belonging to one of the class groups based on data from early time points, and therefore could be accurately classified sooner.

This observation motivates consideration of a sequential classifier, which will sequentially evaluate the subject and decide whether to make a classification at each time point. Such a classifier can be developed by using a sequential neutral zone classifier framework. Different from the traditional classifiers which always assign the subject to one of the class groups, a neutral zone classifier allows to assign a neutral classification (not belonging to any of the class groups) when there is not enough confidence to classify to any of the class groups. At each time point (starting from the first time point), the confidence in classifying the subject to each of the class groups is evaluated. If there is insufficient confidence in making a classification, a neutral classification is made and another measurement at the next time point is collected. This process continues until there is enough confidence of making a classification or the last time point where data can be collected is reached. Using this sequential procedure allows early decisions on subjects which are easier to classify, and also delays decisions on subjects that are difficult to classify. As a result, the proposed sequential classifier reduces the overall cost when the cost of time is taken into account.

The rest of the paper is organized as follows. Section 2 reviews the neutral zone classifier framework. To implement the neutral zone classifier and make it a sequential procedure for classifying longitudinal data, calculating the posterior probability of the subject belonging to one of the class groups at any given time point is a key step. Section 3 describes calculating such probabilities based on a combined logistic regression model and mixed effects model. The overall sequential classification procedure is detailed in Section 4. A simulation study for evaluating the performance of the sequential classifier