

神经网络模型在纵向随访数据分析中的研究进展

张文茜¹, 苏海霞², 孙丽君¹, 张玉海¹

1. 第四军医大学卫生统计学教研室, 陕西 西安 710032; 2. 第四军医大学流行病学教研室

摘要: 流行病学随访研究中会产生大量的纵向数据, 但该类数据的分析一直是统计学的难点。目前大多采用传统的线性混合模型来处理。该模型对数据的分布要求严格, 且假定疾病是线性进展的, 在应用过程中会受到一定限制。近年来, 有研究者提出采用神经网络模型来处理纵向随访数据。本文就神经网络模型在纵向随访数据分析中的研究现状进行探讨, 为纵向随访数据的分析提供一个新思路。

关键词: 纵向数据; 线性混合模型; 神经网络模型

中图分类号: R195.1 文献标识码: A 文章编号: 1006-3110(2017)01-0127-03 DOI: 10.3969/j.issn.1006-3110.2017.01.040

Research progress on neural network model in longitudinal follow-up data analysis

ZHANG Wen-qian*, SU Hai-xia, SUN Li-jun, ZHANG Yu-hai

* Department of Health Statistics, the Fourth Military Medical University, Xi'an, Shaanxi 710032, China

Corresponding author: ZHANG Yu-hai, E-mail: zhyh@fmmu.edu.cn

Abstract: Longitudinal data are generated in the epidemiological follow-up study, but the analysis of such data is always a difficult point in statistics. At present, we mostly use the traditional linear mixed model to analyze such longitudinal data. The model is strict in the distribution of data, moreover, it is assumed that the disease is a linear progression, its application would be subject to some restrictions. Some researchers have proposed using the neural network model to analyze longitudinal follow-up data in recent years. This review discusses the current research status of neural network model in longitudinal follow-up data analysis and provides a new way for the analysis of longitudinal follow-up data.

Key words: Longitudinal data; Linear mixed model; Neural network model

纵向数据 (longitudinal data) 是指对同一研究对象 (受试者、实验动物等) 在不同时间点上重复测量而得到的由横断面数据和时间序列融合在一起的数据^[1], 在医学领域普遍存在, 特别是在慢性疾病的流行病学随访过程中, 是一种典型的数据形式。该数据结合了横断面数据及时间序列数据的双重优点, 既能描述总体的平均增长趋势, 同时能够描述不同个体间的趋势差异。然而, 纵向数据的统计分析却一直是生物医学领域的难点。近年来, 针对纵向数据的统计分析方法得到了国内外广大科研工作者的关注, 已成为生物统计学研究的热点问题之一^[2]。本文将介绍常见的纵向数据分析方法, 并展望纵向数据分析的新进展。

基金项目: 国家自然科学基金 (81573252)

作者简介: 张文茜 (1991-), 女, 山东人, 在读硕士, 研究方向: 神经网络模型。

通信作者: 张玉海 (1974-), 男, 河北人, 博士, 副教授, 研究方向: 纵向数据与神经网络, E-mail: zhyh@fmmu.edu.cn。

目前大多采用传统的线性混合模型 (Linear Mixed Model) 来处理纵向数据。线性混合模型是用于解释变量间存在相关性时对连续性变量的预测建模, 并且同时考虑了两种效应即固定效应与随机效应^[3], 刻画了反应变量和协变量之间的线性关系。纵向数据研究的一个难点是怎样考虑组内相关, 而线性混合模型能较好地解决这个问题, 所以被广泛的应用于纵向数据的分析。该模型由 Harville 提出, Laird & Ware 把它写成标准形式^[4]: $Y_{ij} = X_{ij}\beta + Z_{ij}b_i + \varepsilon_{ij}$, $b_i \sim N_q(0, \sigma^2 D)$, $\varepsilon_{ij} \sim N_{ni}(0, \sigma^2 \Gamma)$, $b_1, b_2, \dots, b_M, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_M$ 是相互独立的。其中 Y_{ij} 是第 i 个个体在第 j 次观测中的测量值, X_{ij} 是 $n_i \times p$ 维的输入变量, β 是 $p \times 1$ 的向量, 代表固定效应。 Z_{ij} 是 $n_i \times q$ 维的设计矩阵, 对应于随机效应 b_i 。

线性混合模型一般要求数据服从正态分布或某种特定的分布, 截距和斜率服从多元正态分布, 并且假定数据是呈现线性趋势^[5]。然而, 慢性疾病的随访数据实际上很难完全符合上述假定。此类随访数据有几个

显著的特点:(1)数据是纵向的,即研究对象在不同时间点上接受多次测量,各测量值之间是相关的,且相关关系形式复杂;(2)数据不平衡,即随访次数不同,且测量时间间隔一般也是不相等的;(3)数据非线性,很多疾病的进展不是线性的,并具有一定的潜伏期。上述特点导致传统的线性混合模型在处理该类数据时存在一定的局限性。近年来,有研究者提出采用神经网络模型来处理纵向随访数据,为这一领域的研究提供了新思路。

1 神经网络模型

人工神经网络(artificial neural network, ANN)是一种模拟生物神经网络的数学模型,由许多神经元组成,各神经元间通过权值相连^[6]。其具有并行性、非线性、容错性以及自适应学习能力等优点,对样本数据是否服从正态分布或某种特定的分布无要求,对变量间是否有相关性无要求^[7],具有很强的非线性问题处理能力。ANN 中伴有反向传播功能的前馈结构——反向传播神经网络(back propagation neural network)是目前发展最成熟的、应用最广泛的一种人工神经网络。典型的神经网络包括三层神经元结构,分别为输入层、隐藏层和输出层,见图 1。

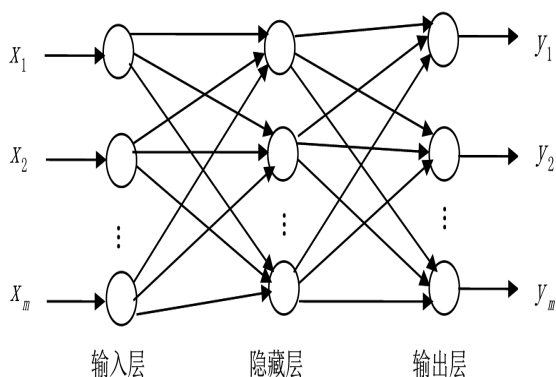


图 1 神经网络模型示意图

人工神经网络在医学诊断、预后、生存分析、临床决策支持等领域中都得到了广泛的应用。Baxt 最早将人工神经网络应用于急性心肌梗死的临床诊断中,并且其结果优于其他分析方法^[8];Da Silva Lopes 等运用次协调神经网络识别阿尔兹海默病^[9];Pappada 等将神经网络应用于胰岛素依赖型糖尿病患者血糖的实时预测^[10]。韩敏等采用主成分分析与神经网络相结合的方法,进行多元变量时间序列的建模和预测研究^[11];陈权等将 BP 神经网络在结直肠癌患者术后生存率预测中的效果与比例风险 Cox 模型、Logistic 回归模型比较,证明 BP 神经网络效果最佳且能有效预测

结直肠癌患者术后生存期^[12];杨同满等对遗传算法以及 BP 神经网络算法的基本原理进行分析,并将具有良好全局搜索能力的遗传算法与能以任意精度逼近非线性函数的神经网络算法相结合,构建基于遗传算法的 BP 神经网络的时间序列预测算法^[13];丁静静等探讨 BP 神经网络在脑胶质瘤患者术后 3 年生存期预测中的应用,为脑胶质瘤患者术后 3 年生存期预测提供了新途径^[14]。

目前,绝大多数的神经网络不能构建纵向数据模型,而充分利用慢性疾病病人不同病情阶段信息的纵向随访数据是构建精确的预测模型的关键。故采用神经网络处理纵向随访数据,从而建立预测模型仍然是一个难点和研究热点。

2 神经网络模型在处理纵向数据中的进展

2.1 纵向数据中时间点的识别 不同于传统的纵向或面板模型,标准的神经网络模型不包括时间相关,因此主要的问题是如何使神经网络模型识别和处理数据中时间或受试者间的关系。对于上述问题,Longhi S 和 Patuelli R^[15-16]等使用了两种方法来获取时间效应或受试者效应。第一种方法是用时间或受试者哑变量和其他协变量作为输入变量。第二种方法是使用一个变量通过文本(字符串)变量的方法识别有关时间点,并通过内部调整文本变量使这种方法成为可能。Tanmay 等^[17]在样本充足的情况下研究神经网络激活函数的变化复杂度对 ANN 的一般作用和特定预测能力的有效性进行研究,并基于统计模型如随机系数模型(线性模型)和非线性混合效应模型(两个有理数模型、一个多项式模型)对神经网络进行模拟。上述研究的模拟模型包括随机参数,但 ANN 参数(权重)是固定的,无法精确地得到估计模型,而混合效应模型能较好地解决这个问题。因此,神经网络模型未来研究的方向是使 ANN 参数能够类似于混合效应模型的随机参数从而精确地得到估计模型。

2.2 混合效应神经网络(mixed effects neural networks, MENN)模型 为解决人工神经网络模型不能处理纵向随访数据的问题,有研究者提出了一种全新的设想——用混合效应神经网络模型来处理纵向数据^[18]。MENN 是一种非参数的神经网络方法,利用线性混合模型的原理,通过创建输入变量的非线性函数与神经网络技术相结合,从而达到既可以处理固定效应和随机效应,又可以处理数据的非线性特征的目的。

MENN 是通过一个输入变量的非线性函数将线性混合模型一般化,可以表示为: $Y_i = f(X_i, \beta) + Z_i b_i + \varepsilon_i, i$

$=1, 2, \dots, M$, 其中 $Y_i = (y_{i1}, y_{i2}, \dots, y_{in_i})$ 是第 i 个病人在第 n_i 次的测量值。 $f(X_i, \beta)$ 是带有输入变量的神经网络部分。 X_i 是 $n_i \times p$ 维的输入变量(协变量), β 是 $p \times 1$ 的向量, 代表固定效应或者是输入变量的人群效应。 Z_i 是 $n_i \times q$ 维的设计矩阵, 对应于随机效应, 或者是个体的特殊效应 b_i 。该模型需构建适当形式的输入变量(协变量)的非线性函数。

从理论上讲, MENN 模型既具有神经网络对数据分布要求不严格的灵活性, 又具有混合效应模型的优势, 对于分析一些慢性疾病纵向随访数据非常合适, 是对传统的线性混合模型的扩展和补充。它具有如下特点: 能够估计随机效应, 从而可以处理重复测量数据; 可以处理不同时间间隔、不同随访次数的非平衡的纵向数据; 神经网络模型非常灵活, 借助于输入变量的非线性函数, 可以根据疾病的特征构建不同的非线性模型, 而慢性疾病的纵向随访数据除了具有一般纵向数据的特征外, 还具有非线性、变异度高等特征, 因此, MENN 非常适合处理这一问题。基于混合效应神经网络模型可在一定程度上解决当前慢性疾病随访数据缺乏合适的分析方法的问题, 该模型可用于建立慢性疾病进展预测模型, 可以为临床医生提供决策支持。

3 讨论与展望

上述神经网络模型为纵向数据的统计学处理提供了一个新思路, 特别是混合效应神经网络模型既具有神经网络对数据分布要求的灵活性, 同时又又可以处理随机效应。通过与输入变量的非线性函数的结合, 就可以分析非线性进展趋势的纵向随访资料, 从而弥补了传统线性混合模型在这一领域的局限性。

但混合效应神经网络模型处于理论探讨阶段尚不完善, 实际应用较少, 还存在诸多问题需要解决, 主要表现在: 模型理论还不完善, 模型的适用性仍需要进一步的验证, 尚无一个完整的方法学体系。Tandon 等^[18]是以老年痴呆症的进展为反 S 形曲线构造的 Sigmoid 函数, 是否适用于其他慢性疾病的纵向随访数据需要进一步的研究; 对于反应变量为离散变量的纵向数据的处理仍然是个难点, 而离散变量在随访数据中非常常见, 影响了该方法的适用性。另外, 缺失值问题在慢性疾病的纵向随访数据分析中也是一个难点, 如何处

理缺失值以及它对 MENN 模型的影响需要进一步的研究。

参考文献

- [1] 庄严. 纵向数据的实验设计及统计分析理论[J]. 数理医药学杂志, 2011, 24(1): 75-77.
- [2] Hedeker D, Gibbons RD. Longitudinal data analysis[M]. New Jersey: John Wiley & Sons, 2006: 105-108.
- [3] 李会民, 王普, 方丽英, 等. 基于混合效应模型的纵向数据建模方法研究[J]. 内蒙古大学学报(自然科学版), 2014, 45(1): 79-83.
- [4] Laird NM, Ware JH. Random-effects models for longitudinal data[J]. Biometrics 1982: 963-974.
- [5] Fishbaugh J, Durrleman S, Piven J, et al. A framework for longitudinal data analysis via shape regression[J]. Proc Spie, 2012, (3): 208-220.
- [6] 张磊, 孔桂兰, 马谢民. 人工神经网络在糖尿病住院费用研究中的应用[J]. 中国医院管理, 2015, 35(1): 61-64.
- [7] Bishop CM. Pattern recognition and machine learning[M]. New York: Springer, 2006: 66-78.
- [8] Baxt WG. Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion[J]. Neural Comput, 1990, 2(4): 480-489.
- [9] Da Silva Lopes HF, Abe JM, Anghinah R. Application of paraconsistent artificial neural networks as a method of aid in the diagnosis of Alzheimer disease[J]. J Med Syst, 2010, 34(6): 1073-1081.
- [10] Pappada SM, Cameron BD, Rosman PM, et al. Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes[J]. Diabetes Technol The, 2011, 13(2): 135-141.
- [11] 韩敏, 席剑辉, 范明明. 神经网络应用于多元变量时间序列的建模研究[J]. 仪器仪表学报, 2006, 19(3): 275-279.
- [12] 陈权. BP 神经网络在结直肠癌预后研究中的应用[D]. 武汉: 华中科技大学, 2011.
- [13] 杨同满, 郭雨. 基于遗传算法的 BP 神经网络时间序列预测算法及其应用[J]. 电脑知识与技术, 2015, 31(1): 160-162.
- [14] 丁静静, 王阿明, 巩萍. BP 神经网络在脑胶质瘤患者术后 3 年生存期预测中的应用[J]. 徐州医学院学报, 2015, 35(6): 415-417.
- [15] Longhi S, Nijkamp P, Maierhofer E. Neural network modeling as a tool for forecasting regional employment patterns[J]. Intern Reg Sci Rev, 2005, 28(3): 330-346.
- [16] Patuelli R, Reggiani A, Nijkamp P, et al. New neural network methods for forecasting regional employment: an analysis of German labour markets[J]. Spatial Econ Anal, 2006, 1(1): 7-30.
- [17] Maity TK, Pal AK. Subject specific treatment to neural networks for repeated measures analysis[J]. Proceed Internl MultiConf Eng Comp Scient, 2013, 1(1): 60-65.
- [18] Tandon R, Adak S, Kaye JA. Neural networks for longitudinal studies in Alzheimer's disease[J]. Artif Intell Med, 2006, 36(3): 245-255.

收稿日期: 2016-07-05