基于逻辑回归模型的缺血性脑卒中发病率预测研究*

李 鹏 闵 慧

（1 湖南中医药大学信息科学与工程学院 长沙 410208（湖南信息职业技术学院软件学院 长沙 410200）
2 中南大学湘雅三医院 长沙 410006
3 医学信息研究湖南省普通高等学校重点实验室
（中南大学） 长沙 410006）

瞿昊宇 罗爱静

（湖南中医药大学信息科学与工程学院 长沙 410208）（中南大学湘雅三医院 长沙 410006）
（医学信息研究湖南省普通高等学校重点实验室
（中南大学） 长沙 410006）

〔摘要〕 介绍基于逻辑回归模型的缺血性脑卒中发病率预测方法及流程，包括收集和清洗数据、构建大数据平台、提取预测特征、构建基于逻辑回归的模型等。通过仿真实验验证该方法的有效性，为脑卒中数据分析、疾病预防提供技术支持。

〔关键词〕 缺血性脑卒中；数据清洗；特征提取；逻辑回归；小批量梯度下降法；预测精度
〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.06.006

Study on the Prediction of the Incidence of Ischemic Stroke Based on Logistic Regression Model *LI Peng, 1School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, 2The Third Xiangya Hospital of Central South University, Changsha 410006, 3Key Laboratory of Medical Information Research (CSU), College of Hunan Province, Changsha 410006, China; MIN Hui, Software Institute, Hunan College of Information, Changsha 410200, China; QU Haoyu, School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China; LUO Aijing, 1The Third Xiangya Hospital of Central South University, Changsha 410006,*

〔修回日期〕 2019－12－27
〔作者简介〕 李鹏，博士，讲师，发表论文25 篇；通讯作者：闵慧，硕士，讲师。
〔基金项目〕 国家社会科学基金重点项目“网络健康信息资源聚合与精准信息服务研究”（项目编号：17AZD037）；国家重点研发计划“中医智能舌诊系统及数据平台研发与应用”（项目编号：2017YFC1703306）；湖南省自然科学基金青年项目“无线传感网中基于压缩感知的数据收集关键技术研究”（项目编号：2019JJ50453）；湖南省自然科学基金面上项目“基于‘法－方－药’网络机器学习的中医治疗银屑病复方功效预测研究”（项目编号：2018JJ2301）；湖南省科技厅重点项目“基于大数据的中西医结合防治脑梗死创新技术研究与推广应用”（项目编号：2017SK2111）；湖南中医药大学开放基金项目“面向动态蛋白质网络的功能模块挖掘方法研究”（项目编号：2018JK02）。

2Key Laboratory of Medical Information Research (CSU), College of Hunan Province, Changsha 410006, China

〔**Abstract**〕 The paper introduces the prediction methods and process of the incidence of ischemic stroke based on logistic regression model, including collecting and cleaning data, building big data platform, extracting prediction features, building a logistic regression – based model, etc., and verifies the effectiveness of this method through simulation experiment, which provides technical support for data analysis and disease prevention of stroke.

〔**Keywords**〕 ischemic stroke; data cleaning; feature extraction; logistic regression; mini – batch gradient descent method; prediction accuracy

1 引言

缺血性脑卒中是指由于脑的供血动脉（颈动脉和椎动脉）狭窄或闭塞、脑供血不足导致的脑组织坏死的总称。近年来缺血性脑卒中[1]已经成为危害人类健康和生命安全的重大疾病，如何有效地对缺血性脑卒中发病率进行预测，识别可能导致缺血性脑卒中疾病的高危因素，提高高危患者风险意识，具有十分重要的意义[2]。目前临床上用于脑卒中筛查或预测复发的相关方法较多，例如汪仁等[3]采用全国脑卒中筛查数据作为训练和测试数据，构建一种基于决策树的脑卒中分级预测方法。朱千里[4]从脑卒中致病原因（是否存在心房颤动）出发，采用人工神经网络对脑卒中发病率进行预测，预测结果可用于指导脑卒中患者的个性化治疗。陈莉平等[5]根据收集的脑卒中数据，构建脑卒中大数据应用平台，开发基于 AdaBoost 的脑卒中复发预测模型对脑卒中初患人群进行复发风险预测。本文针对现有方法的不足提出一种基于逻辑回归模型的缺血性脑卒中发病率预测方法。通过收集和清洗数据、提取面向缺血性脑卒中预测特征、构建基于逻辑回归的模型等过程来实现缺血性脑卒中发病率的预测，最后通过仿真实验验证方法的有效性。

2 基于逻辑回归的缺血性脑卒中发病率预测

2.1 具体流程（图1）

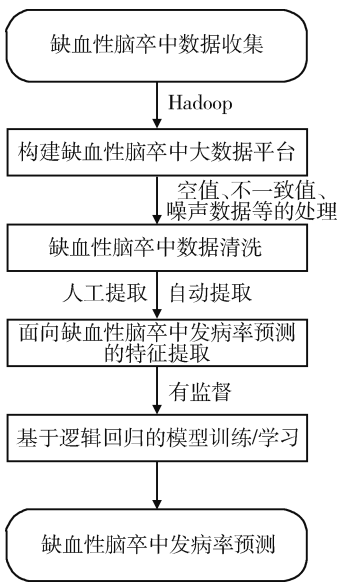

图1 基于逻辑回归的脑卒中预测流程

2.2 数据收集和平台构建

首先利用数据接入及导入工具对分散在基地医疗机构、社区卫生中心、保健机构、体检机构和三甲医院等各级机构中的患者信息进行采集和集成，最终形成缺血性脑卒中患者病历信息库。采集内容涉及患者个人信息、既往史、家族史、住院诊疗数据、阶段性随访数据、体检数据等。在数据收集的基础上，采用 Hadoop[6]作为基本的分布式执行架构，在该架构上配置 Python 与 Spark 等分析工具，构建集脑卒中患者数据采集、存储、分析、模型学习、疾病诊治等功能一体化的大数据平台，见图2。

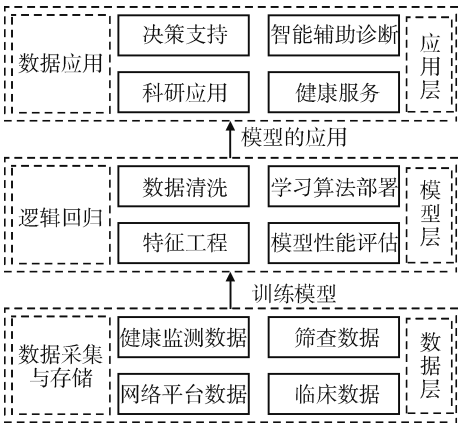

图2 缺血性脑卒中大数据平台

2.3 数据清洗

脑卒中管理数据来源广泛，形式多样，涉及种类很多，且由于受到筛查对象主观性、时间限制、信息获取成本高等因素影响，收集到的脑卒中大数据经常存在空值、不一致、噪声数据等。因此需要对这些数据进行预处理以提高后续预测方法的准确性。其中空值数据对于算法的影响很大，采用删除包含空值的记录、自动和手工补全缺失值等方法处理；对于不一致数据，则在分析产生原因的基础上利用各种变换、格式化、汇总分解函数实现数据清洗；对于噪声数据，采用分箱、计算机与人工检查相结合和聚类3 种方法处理。

2.4 特征提取

2.4.1 概述 实证研究和相关统计表明[7]目前影响缺血性脑卒中发病的高危因素包括：年龄、遗传、高血压、高血脂、高血糖、心脏病、不良饮食、缺乏运动、吸烟、酗酒。从预处理后的脑卒中数据集中提取上述10 种因素作为特征来进行模型训练。考虑到基于逻辑回归模型的输入一定是数值类型，而提取的10 个特征中大部分是字符串类型，需要将字符串类型转换成数值类型，如向量、矩阵或张量形式。一般而言常见特征可以分为类别和数值型特征两大类。其中对于类别特征，使用独热编码[8]技术将其转换为数值类型后再作为模型的输入。对于数值型特征，直接对其进行特征归一化后将其作为模型的输入。

2.4.2 特征编码 将分类特征表示为二进制向量，又称一位有效编码，其方法是使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都有独立的寄存器位，在任意时间其中只有一位有效。提取的10 个特征中遗传、心脏病、不良饮食、缺乏运动、吸烟和酗酒6 个特征属于类别特征，需要对其进行独热编码后再作为模型的输入。编码过程，见图3。

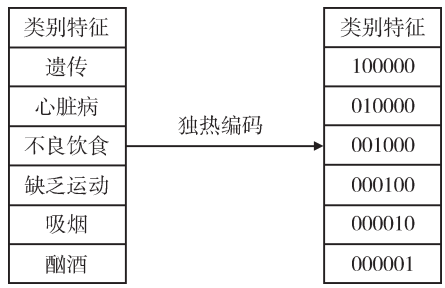

图3 独热编码

2.4.3 特征归一化 提取的10 个特征中年龄、高血压、高血脂和高血糖4 个特征属于数值类特征，对其进行特征归一化处理，以提高模型精度和训练过程中算法的收敛速度。采用两种常见的特征归一化方法：线性归一化和标准差标准化。其中线性归一化是指将特征值范围映射到［0，1］区间，见公式1；标准差标准化的方法是指将特征值映射到均值为0、标准差为1 的正态分布，见公式2。

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

$$x_{new} = \frac{x - mean(x)}{std(x)} \quad (2)$$

其中 $\min(x)$ 指 x 的最小值，$\max(x)$ 指 x 的最大值，$mean(x)$ 指 x 的平均值，$std(x)$ 指 x 的标准差。以10 个样本的年龄特征为例，根据上述公式对其进行特征归一化的结果，见图4。

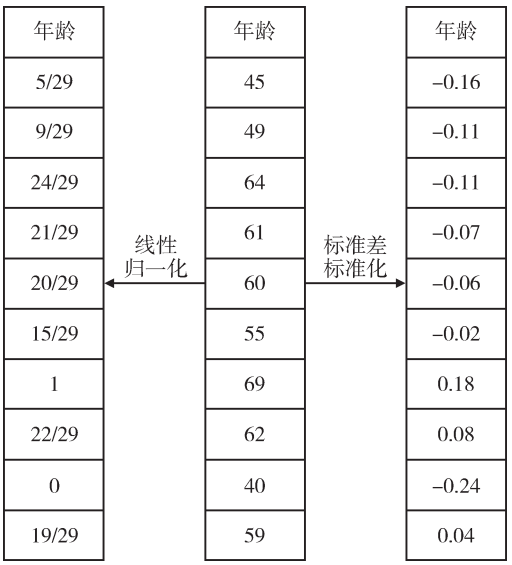

图4 年龄特征的归一化

2.5 基于逻辑回归的模型训练

2.5.1 概述 逻辑回归又称为 logistic 回归分析[9]，是一种广义的线性回归分析模型，常用于数据挖掘、疾病自动诊断、经济预测等领域。以脑卒中病情分析为例，选择两组人群，一组是脑卒中患者组，一组是非脑卒中患者组，两组人群必定具有不同的体征与生活方式等。因变量为是否患上脑卒中，值为“是”或“否”；自变量为上述影响脑卒中发病的 10 大特征。自变量可以是连续或是分类的，通过 logistic 回归分析可以得到自变量最优权重，从而准确预测不同人群患脑卒中的可能性。

2.5.2 确定目标函数 首先基于逻辑回归模型将缺血性脑卒中发病率预测问题采用以下数学表达式进行建模：

$$\begin{cases} y = \dfrac{1}{1+e^{-z}} \\ z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_{10} x_{10} \end{cases} \tag{3}$$

其中 y 指待观测个体患上缺血性脑卒中的概率，是一个 Sigmoid 函数[10]，采用该函数的意义在于不管影响脑卒中的因素有多少，最终得到的是一个关于缺血性脑卒中发病率的取值在［0，1］之间的概率。$x_1, x_2, \ldots, x_{10}$ 指影响缺血性脑卒中发病的 10 大特征，θ 是权重参数。为使预测结果与真实结果的误差最小化，采用最小化均方误差[11]作为逻辑回归的损失函数，从而得到本研究预测问题的优化目标为：

$$\min_{\theta} y = \frac{1}{2m}\sum_{i=1}^{m}(y^{(i)} - y_\theta(x^{(i)}))^2 \tag{4}$$

其中 m 是样本的规模；$y_\theta(x^{(i)})$ 是对第 i 个样本进行训练得到的预测结果；$y^{(i)}$ 是第 i 个样本的真实结果（标签）。要构建准确的缺血性脑卒中发病率预测模型，即要求解得到式（4）中的参数 θ 的最优值。

2.5.3 模型求解 为求解式（4）的优化问题，常采用最小二乘法，将求解式（4）的优化问题转化为求函数极值问题，但这种做法并不适合计算机。为此采用小批量梯度下降法（Mini - batch Gradient Descent，MBGD）[12]进行模型求解。该方法训练过程比较快，且能保证最终参数训练的准确率。特点是每次训练迭代在训练集中随机采样 M 个样本，其数学表达式为：

$$\begin{cases} \dfrac{\partial\, \gamma}{\partial\, \theta_j} = -\dfrac{1}{M}\sum_{i=1}^{M}(y^{(i)} - y_\theta(x^{(i)}))x_j^{(i)} \\ \theta_j := \theta_j + \eta\dfrac{1}{M}\sum_{i=1}^{M}(y^{(i)} - y_\theta(x^{(i)}))x_j^{(i)} \end{cases} \tag{5}$$

其中 $x_j^{(i)}$ 是第 i 次迭代中选定样本的第 j 个输入分量；η 是学习率。根据式（3）进行反复迭代来更新参数 θ，直到两次迭代之间的误差低于某一固定阈值时更新结束。此时得到的 θ 值即为最优参数，将其代入式（3）中得到缺血性脑卒中发病率预测模型。只要将任何一名疑似患者的 10 大特征数据（$x_1, x_2, \ldots, x_{10}$）输入公式（3）中就可以准确地预测其患缺血性脑卒中的概率，为临床医生提供诊治决策依据，例如预测概率超过 0.6 则可认定该疑似患者极有可能发生缺血性脑卒中，应该重点监护。

3 实验

以获取到的全国 2012 - 2018 年缺血性脑卒中院外筛查数据作为研究对象，覆盖全国 31 个省市自治区总计 454 个筛查点，随机选定城乡社区的 40 岁及以上常驻人群进行社区整群抽样获得数据。截至目前累计收集并存储近 700 万人的院外筛查档案。

本文从这些档案数据中随机抽样 50 000 条档案作为数据集。其中 70% 数据集为训练集，30% 数据集为测试集。将本文提出的预测方法与决策树算法[3]、人工神经网络算法[4]和 AdaBoost 算法[5]进行性能对比，采用查准率来评价各种算法性能。不同方法查准率比较结果，见图 5。可以看出随着数据规模的增加，4 种方法的预测精度都有不同程度的上升。总的来看，本文方法的预测精度要略高于人工神经网络算法，比决策树算法和 AdaBoost 算法的预测精度分别高出约 18.8% 和 21.7%。分析原因可知：一是本文预测方法在建模过程中采用多种技术对缺血性脑卒中原始大数据进行清洗，并对影响脑卒中的高危因素进行分析和特征提取，将噪声数据对模型的影响降到最低；二是对每个特征进行特征编码或归一化的分类处理，提高特征对于模型的吻合度；三是采用小批量梯度下降法在降低训练时间的同时进一步保证预测准确性。

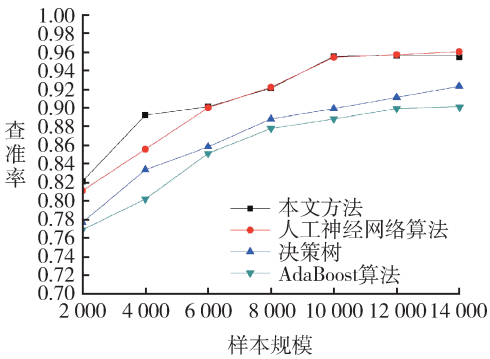

图 5 不同方法查准率比较

4 结语

本文提出一种基于逻辑回归模型的缺血性脑卒中发病率预测方法并通过实验验证其有效性。下一步工作中将采用深度学习技术自动提取影响缺血性脑卒中发病的重要因素，设计一种基于图卷积神经网络的缺血性脑卒中发病率预测方法，为医生智能诊疗提供更好的技术支持。

参考文献

1 郭健，刘远立，关天嘉，等. 健康相关行为与高血压人群脑卒中发病风险的关联［J］. 中华预防医学杂志，2019，53（2）：223－228.

2 张晓莉，唐朝正，贾杰. 中西医治疗脑卒中后肩手综合征现状分析［J］. 中国康复医学杂志，2015，30（3）：294－298.

3 汪仁，边迪，王树奇，等. 决策树算法在脑卒中危险分级预测中的应用［J］. 中国疗养医学，2019，28（3）：233－236.

4 朱千里. 人工智能在脑卒中患者心房颤动预测中的应用［J］. 医院管理论坛，2019，36（7）：30－33.

5 陈莉平，宋立冉. 基于大数据的脑卒中复发预测模型的构建［J］. 物联网技术，2019，9（6）：50－54.

6 Tariq H，Al－Sahaf H，Welch I. Modelling and Prediction of Resource Utilization of Hadoop Clusters：a machine learning approach［C］. Paris：Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing. ACM Press，2019：93－100.

7 孙慧君，鲁健萍，张洁，等. 脑卒中健康素养的研究进展［J］. 中国全科医学，2019，22（36）：4409－4414.

8 Rodríguez P，Bautista M A，Gonzalez J，et al. Beyond One－hot Encoding：lower dimensional target embedding［C］. Chongqing：The 3rd IEEE International Conference on Image and Vision Computing. IEEE Press，2018：21－31.

9 朱燕波，王琦，吴承玉，等. 18805 例中国成年人中医体质类型与超重和肥胖关系的 Logistic 回归分析［J］. Journal of Integrative Medicine，2018，8（11）：1023－1035.

10 Iliev A，Kyurkchiev N，Markov S. On the Approximation of the Step Function by Some Sigmoid Functions［J］. Mathematics and Computers in Simulation，2017（133）：223－234.

11 隋昊，覃高峰，崔祥波，等. 基于误差均值与方差最小化的鲁棒 TS 模糊建模方法［J］. 浙江大学学报（工学版），2019，53（2）：382－387.

12 Liu J，Takáč M. Projected Semi－stochastic Gradient Descent Method with Mini－batch Scheme under Weak Strong Convexity Assumption［C］. Bethlehem：Modeling and Optimization：theory and applications，2016：95－117.