

# 应用分类模型研究迟发性颅脑损伤的影响因素

史宝鹏,段 迅,孔广黔,吴 云

(贵州大学 计算机科学与技术学院,贵州 贵阳 550025)

**摘 要:**迟发性颅脑损伤是危害人类健康及生命的常见疾病之一。文中使用 SPSS 统计分析软件根据已有的患者信息进行分析,并使用模型联合应用技术,以逻辑回归为主模型给出明确的回归方程;以决策树模型为辅助模型探索变量间的交互作用;用探索结果指导逻辑回归的建模,使得模型更加准确。实验结果表明,激素是预防迟发性颅脑损伤作用最大的因素;舒张压和血小板对迟发性颅脑损伤的发生也有较大影响;同时,舒张压和血小板交互作用对迟发性颅脑损伤的发生也有一定影响。这一研究发现能更快更好地找出导致迟发性颅脑损伤的主要原因,辅助医生对患者是否发生迟发性颅脑损伤做出判断并做出更为精准的诊疗方案,降低患者发生迟发性颅脑损伤的概率。

**关键词:**数据挖掘;分类模型;逻辑回归;决策树;医疗

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2018)03-0201-04

doi:10.3969/j.issn.1673-629X.2018.03.043

## Study on Influencing Factors of Delayed Craniocerebral Brain Injury by Classification Model

SHI Bao-peng, DUAN Xun, KONG Guang-qian, WU Yun

(School of Computer Science & Technology, Guizhou University, Guiyang 550025, China)

**Abstract:** The delayed brain injury is one of the common diseases of endangering human health and life. According to SPSS statistical analysis software to analyze the existing patient information, we use the model joint application technology to give the regression equation with the logistic regression as the main model, explore the interaction between variables with decision tree model as the auxiliary and guide the logistic regression modeling with the exploration results, which makes the model more accurate. The experiments show that the hormone is the most important factor in preventing delayed brain injury. The diastolic blood pressure and the platelet have a great influence on the occurrence of delayed brain injury, and the interaction of them do so at the same time. The study can find the main cause of delayed brain injury faster and better, which assists the doctors to determine whether the patients have a delayed brain injury and to make a more accurate diagnosis and treatment program for reduction of the probability of patients with delayed traumatic brain injury.

**Key words:** data mining; classification model; logistic regression; decision tree; medical treatment

## 0 引 言

近年来,国内医疗信息化事业飞速发展。据统计,国内 80% 医疗机构采用 HIS 系统(医疗信息系统)办公,每天对大量的医疗、财务、药品及物资等信息进行管理,但对数据处理仅限于简单的录入及查询,在医疗数据分析和辅助决策方面发展较为缓慢<sup>[1-3]</sup>。如何针对临床诊疗信息、财务信息等海量数据进行有效模式的挖掘,通过信息的分类及分析,发现医疗业务和管理中的潜在问题,辅助医生及行政人员做出正确决策,提

高医疗机构的医疗及管理水平,是医疗机构急待解决的问题<sup>[4-5]</sup>。基于此,文中应用数据挖掘模型对迟发性颅脑损伤患者诊疗信息进行分析,找出引发迟发性颅脑损伤的主要影响因素,辅助医生做出诊疗决策,减少患者的发病率。

## 1 背 景

### 1.1 数据挖掘概述

数据挖掘是从大量、不完整、有噪音、看似无关的

收稿日期:2017-03-25

修回日期:2017-07-30

网络出版时间:2017-12-04

基金项目:贵州省科技计划项目;基层远程诊断服务平台云数据中心(黔科中引地[2016]4008号)

作者简介:史宝鹏(1990-),男,硕士研究生,研究方向为云平台、数据挖掘;段 迅,通讯作者,博士,副教授,研究方向为云计算、数据挖掘和传感网络;孔广黔,博士,副教授,CCF 会员(06800M),研究方向为云计算和计算机网络;吴 云,博士,副教授,研究方向为分布式计算、数据挖掘及其应用。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20171204.1647.020.html>

实际应用数据中,挖掘出令人感兴趣的、有价值的、隐含的、事前未知的模式或知识。模式或知识的发现过程一般包括数据清理、数据集成、数据转换、数据挖掘、模式评估及知识表示<sup>[6-9]</sup>。数据挖掘技术能自动化地分析数据源中的数据,并做出归纳、推理。通过数据挖掘,有价值的模式或规则从数据源中被抽象并展示出来。数据挖掘是信息技术发展的必然结果,主要涉及数据库、统计学及机器学习等学科。其主要有关联分析、分类、聚类及预测四大功能。

在医学领域中,数据挖掘技术有其自身的优势。在医学领域中收集的数据大多是真实可靠的遗漏数据和噪音数据比例较少的结构化数据,不但减轻了数据挖掘中数据清理、数据集成和数据转换的工作量,并且使得医疗数据具有较强的稳定性,这些因素有益于数据挖掘模式和知识的维护和质量保证<sup>[10-13]</sup>。

### 1.2 分类模型在迟发性颅脑损伤研究中的意义

迟发性颅脑损伤是危害人类健康的常见疾病之一。由于车祸、高处坠落及暴力打击等外界因素导致颅脑损伤,在初期并未发现颅内血肿等颅脑疾病,但经过一段时间后再次检查时发现颅内血肿等脑部损伤,这种脑损伤往往会在人们疏忽时发病,导致较高的致残率和死亡率。迟发性颅脑损伤的发病率正在逐年上升,这种病症也日益受到医学工作者的重视<sup>[14-15]</sup>。

文中根据某省医院的脑外科医生收集的该科室在3年间急救后治疗的201例脑外伤病例,使用分类模型从中分析出导致急救后迟发性颅脑损伤的主要影响因素,确定是否发生迟发性颅脑损伤变量受到哪些影响因素的作用,以辅助医生做出合理决策及诊疗方案,有效提高治愈率,降低患者的致残率和死亡率。

## 2 关键技术

文中使用两种分类模型对迟发性颅脑损伤的主要影响因素进行分析,属于分类问题中因变量的影响因素的发现与确认。因变量(是否发生迟发性颅脑损伤)为二分类变量,候选变量不是单一变量,需要多因素建模,因此选用基于线性模型发展而来的逻辑回归为主分析模型。逻辑回归可以满足对分类因变量进行多变量建模的需求,模型中也可以同时纳入连续型自变量和分类的自变量。逻辑回归无法做变量间的劣效性检验,在分类数据的多变量模型中确定变量间交互作用时工作十分繁琐。因此以决策树模型作为辅助模型,探索变量间的交互作用,使得实验完整、实验结果更加准确可信。

### 2.1 逻辑回归模型

因变量 $Y$ 为一个二值变量,即 $Y=0$ 或 $Y=1$ ;自变量为 $X_1, X_2, \dots, X_m$ ;  $P$ 表示在 $m$ 个自变量的作用下 $Y$

发生的概率,由式(1)所示:

$$P = 1 / (1 + e^{-Z}) \quad (1)$$

其中, $P$ 的取值范围为 $(0,1)$ 。

统计量 $Z$ 为:

$$Z = \beta_0 + \beta_1 X_1 + \dots + \beta_m X_m \quad (2)$$

其中, $\beta_0$ 为常数项,表示所有影响因素均为0时个体发生概率与不发生概率之比的自然对数的变化值; $\beta_1, \beta_2, \dots, \beta_m$ 为回归系数,表示某个因素 $X_i$ 改变一个单位时个体发生概率与不发生概率之比的自然对数的变化值; $Z$ 的取值范围为 $(-\infty, +\infty)$ 。

逻辑回归中最重要的两步是参数估计和变量选择。逻辑回归采用最大似然估计的方法估计回归系数 $\beta_1, \beta_2, \dots, \beta_m$ ,同时得到回归系数的标准误差 $S_{\beta}$ 。所有样本预测值与真实值一致的概率 $\iota(\beta)$ 最大时的回归系数即为所求。其中 $P(y_i)$ 为单个样本预测值与真实值一致的概率:

$$P(y_i) = X_i^{y_i} (1 - X_i)^{(1-y_i)} \quad (3)$$

所有样本预测值与真实值一致的概率为:

$$\iota(\beta) = \prod P(y_i) \quad (4)$$

首先对式(4)两边取对数,然后对 $\beta_i$ 求偏导,最后利用牛顿迭代法求得回归系数的值。

当影响因素过多时,需挑选出与事件发生确实有关系或是关系更密切的影响因素,建立更加稳固的回归模型。筛选变量的方法有前进法、后退法、逐步法、似然比检验法和Wald检验法等,根据变量的统计量意义筛选或剔除变量。

逻辑回归不但泛化能力强、精准度高,而且能精确控制用户数量。但是逻辑回归对数据要求较高,不能处理复杂的用户特征及共线性的问题。医疗数据大多为完整的结构化数据,逻辑回归在其上的应用有着独特的优势。

### 2.2 决策树模型

决策树从根节点开始,每一层节点依照某一属性值向下分裂子节点,待分类的实例在每一节点处比较该实例各个属性的信息增益,根据信息增益最大的属性向相应的子节点扩展,这一过程在到达决策树的叶子节点时结束。

划分前信息量:设数据集 $D$ 为类标记的元组训练集,假设类标号属性具有 $M$ 个不同的值,定义 $m$ 个不同的类 $C_i (i=1,2,\dots,m)$ 。对 $D$ 中的元组分类所需的期望信息量为:

$$\text{Info}(D) = - \sum P_i * \log_2(P_i) \quad (5)$$

划分后信息量:假设属性 $A$ 具有 $V$ 个不同的离散属性值,可使用属性 $A$ 把数据集 $D$ 划分为 $v$ 个子集 $\{D_1, D_2, \dots, D_v\}$ ,设子集 $D_j$ 中全部的记录数在 $A$ 上具

有相同的值  $\alpha_j$ 。基于按  $A$  划分对  $D$  的元组分类所需要的期望信息量为:

$$\text{Info}_A(D) = - \sum (D_j/D) * \text{Info}(D_j)$$

(6)

信息增益为原来的信息量(基于类比例)与新的信息量(对  $A$  划分后)之间的差:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

(7)

决策树模型可以生成易被理解的规则集,业务解释性较好,具有较好的健壮性,能够很好地处理非线性关系。但是当类别过多时误判率会明显增加,且泛化能力较差。

### 3 迟发性颅脑损伤信息的挖掘流程

#### 3.1 原始变量

通过对收集到的迟发性颅脑损伤的医疗数据进行整理后,得到用于研究的变量主要有 ID、性别、年龄、收缩压、舒张压、血小板、脑挫伤、手术、中线移位、脑肿胀、意识程度、止血药、激素和脱水剂。

#### 3.2 数据理解

(1)单变量描述/数据变换。

在变量描述时需要对连续型变量进行描述,其中

表2 分类变量检验

|     |   | 性别  |    | 脑挫伤 |    | 手术  |    | 中线移位 |    | 脑肿胀 |    | 意识程度 |    |    | 止血药 |     | 激素 |     | 脱水机 |     |
|-----|---|-----|----|-----|----|-----|----|------|----|-----|----|------|----|----|-----|-----|----|-----|-----|-----|
|     |   | 1   | 2  | 0   | 1  | 0   | 1  | 0    | 1  | 0   | 1  | 1    | 2  | 3  | 0   | 1   | 0  | 1   | 0   | 1   |
| 迟发  | 0 | 116 | 40 | 57  | 99 | 109 | 47 | 115  | 41 | 124 | 32 | 82   | 33 | 41 | 3   | 153 | 1  | 155 | 56  | 100 |
| 脑损伤 | 1 | 31  | 14 | 21  | 24 | 27  | 18 | 26   | 19 | 35  | 10 | 11   | 24 | 10 | 3   | 42  | 6  | 39  | 23  | 22  |

对每个连续变量多做一个 T 检验,实验结果如表 3 和表 4 所示。可见患者和非患者的收缩压、舒张压及血小板是有区别的,而患者和非患者的年龄是没有区别的。

表3 变量分组均值

| 变量     | 迟发脑损伤 |      |
|--------|-------|------|
|        | 0     | 1    |
| 年龄     | 40    | 42   |
| 收缩压    | 16    | 13   |
| 舒张压    | 10    | 7    |
| 血小板对数值 | 4.95  | 4.44 |

表4 列均值的比较

| 变量     | 迟发脑损伤 |   |
|--------|-------|---|
|        | 0     | 1 |
| 年龄     |       |   |
| 收缩压    | B     |   |
| 舒张压    | B     |   |
| 血小板对数值 | B     |   |

#### 3.3 逻辑回归建模

将迟发性脑损伤作为因变量,将所有经过预筛选后需要进一步分析的变量选为协变量。

模型中存在无效变量需要化简模型,化简模型,剔

除年龄和血小板为连续变量,因此描述结果如表 1 所示。

表1 连续变量的描述

| 变量    | N   | 极小值 | 极大值 | 均值     | 标准差   |
|-------|-----|-----|-----|--------|-------|
| 年龄    | 201 | 12  | 86  | 40.75  | 18.92 |
| 血小板   | 201 | 51  | 423 | 137.15 | 63.57 |
| 有效的 N | 201 |     |     |        |       |

从表 1 可见,血小板极小值为 51,极大值为 423,范围过大,可能有极端值或是偏态导致此问题的发生,需要对血小板进行进一步描述。

变量血小板为偏态分布,没有发生迟发脑损伤的血小板水平明显偏高,发生迟发脑损伤的血小板水平明显偏低,可以看出血小板水平可能是迟发脑损伤的影响因素。由于血小板是偏态分布且是自变量,转换后在临床上解释更为合理,因此需要将其转换为变量:ln 血小板(即血小板的自然对数值)。转换后血小板分布较为对称。

(2)单变量的分析及变量筛选。

这个过程主要用表一次性把分类和连续变量与因变量的联系表示出来。对每个分类变量多做一个卡方检验,检验各变量与迟发性脑损伤是否有关联,结果如表 2 所示。

除  $P$  值最大的变量收缩压和止血药,同时要考虑变量间共线性的问题。对模型进行比较,比较似然比检验值,结果如表 5 所示。似然值表示模型对数据的解释程度,最理想的情况是该值应无限接近于 0,该值越大表示对数据的解释性越差。剔除变量后该值的-2 对数似然值为 68.147,提出变量前该值的-2 对数似然值为 68.015。剔除变量后该值上升 0.132,由此说明剔除变量为无关变量。

表5 似然值比较

|     | -2 对数似然值            | Cox&Snell R 方 | NagelKerke R 方 |
|-----|---------------------|---------------|----------------|
| 变量前 | 68.015 <sup>a</sup> | .516          | .788           |
| 变量后 | 68.147 <sup>a</sup> | .515          | .787           |

接下来依次剔除脑肿胀、脑挫伤等变量,最终剩下舒张压、激素及 ln 血小板三个变量。该模型似然值为 72.987,较之前有明显上升,说明该模型更优秀。

各变量的解释说明:舒张压每增加一个单位,相应的个体发生脑损伤的概率就降低 29.8%,不打激素的患者发生脑损伤的概率是打激素的患者发生脑损伤的概率的 21 772.131 倍,ln 血小板每增加一个单位发生脑损伤的概率就降低 0.004。由此可见,激素是可控



的重要因素,也是最核心的抢救措施。

该模型存在问题:实验中被剔除的变量在主效应中无效但交互项有意义,需考虑被剔除的没有统计学意义的变量间是否存在交互项需要保留。由于变量及其组合过多,构成的模型会发生混乱。对于连续型变量需对其做标准正态变换然后再添加至候选变量,还需手工构建代表相应交互作用的新变量。高阶交互项需要劣效性检验,而逻辑回归中并无劣效性检验。

### 3.4 决策树模型

将总研究人群通过某些特征(自变量取值)分成数个相对同质的亚人群,使得每个亚人群内部的因变量取值高度一致,而不同亚人群间的因变量取值差异较大。树模型结构可以解决交互项及影响因素的发

现,可用于分类变量或连续变量的分类。树模型会在所有候选变量进行筛选,按照重要性的大小依次挑选出自变量进入模型,在处理大量自变量的分析问题中性能较好。树模型均为非参数方法,没有太多的使用条件限制,应用范围广,适用于复杂的联系分析。但不能对影响因素的作用大小进行精确的定量描述,对于因变量和自变量间是线性关联、无交互作用时效果可能不是很理想。样本量需要充足才能保证逐层细分后单元格内仍有充足的样本数。

使用决策树模型对样本进行分类,可以看出血小板与舒张压存在交互项。在逻辑回归中添加血小板与舒张压的交互项,用树模型解决交互项的搜索和确认的问题,结果如表 6 所示。

表 6 最终模型

| 变量           | B       | S. E   | Wals   | df | Sig. | Exp(B)     |
|--------------|---------|--------|--------|----|------|------------|
| 舒张压          | -11.348 | 4.010  | 8.008  | 1  | .005 | .000       |
| 激素           | -8.573  | 2.014  | 18.119 | 1  | .000 | .000       |
| ln 血小板       | -26.747 | 9.166  | 8.515  | 1  | .004 | .000       |
| ln 血小板 & 舒张压 | 2.221   | .871   | 6.496  | 1  | .011 | 9.214      |
| 常量           | 140.136 | 43.172 | 10.537 | 1  | .001 | 7.252E+060 |

对预防迟发性脑损伤作用最大的指标是激素,结果显示使用激素会使迟发性脑损伤的发生风险降至原来的 2 万分之一(即  $e^{-9.988}$ );舒张压和血小板对数值也有一定作用,但其作用明显弱于激素;分析发现舒张压和血小板对数值间存在协同的交互作用。

## 4 结束语

针对实际收集的脑外伤患者数据,应用数据挖掘中的模型联合应用技术,以逻辑回归为主模型,给出明确的回归方程,清晰易懂的结果解释,但是在进行交互项的查找和验证方面效果欠缺。基于此,应用决策树模型做扩展性的探索,发现变量间潜在的交互作用,用结果指导逻辑回归的建模。最终确定舒张压、激素、血小板及激素与血小板交互项为急救后迟发性颅脑损伤的主要影响因素。文中不足之处在于样本量不充分,在辅助模型决策树模型中叶子节点样本不充足,后期会对大量样本进行分析,不断提高实验结果的准确性。

### 参考文献:

[1] 沈佳,杨渭林,裴申忠,等.重型颅脑损伤患者行开颅手术后发生迟发性颅内血肿的危险因素分析[J].中国全科医学,2014,17(33):3997-3999.

[2] 张丽娟,李舟军.分类方法的新发展:研究综述[J].计算机科学,2006,33(10):11-15.

[3] 罗可,林睦纲,郗东妹.数据挖掘中分类算法综述[J].计算机工程,2005,31(1):3-5.

[4] ROBERT N, JOHN E, GARY M. Handbook of statistical a-

nalysis and data mining applications[M]. [s. l.]: Academic Press, 2009.

[5] 熊平.数据挖掘算法与Clementine实践[M].北京:清华大学出版社,2011:44-60.

[6] 徐鹏,林森.基于C4.5决策树的流量分类方法[J].软件学报,2009,20(10):2692-2704.

[7] CHAO C M, YU Y W, CHENG B W, et al. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree[J]. Journal of Medical Systems, 2014, 38(10):106.

[8] KANTARDZIC M. 数据挖掘:概念、模型、方法和算法[M].北京:清华大学出版社,2003.

[9] WANG Yaonan, YUAN Xiaofang. SVM approximate-based internal model control strategy[J]. Acta Automatica Sinica, 2008, 34(2):172-179.

[10] 韩松来,张辉,周华平.基于关联度函数的决策树分类算法[J].计算机应用,2005,25(11):2655-2657.

[11] 王光宏,蒋平.数据挖掘综述[J].同济大学学报:自然科学版,2004,32(2):246-252.

[12] 马秀红,宋建社,董晨飞.数据挖掘中决策树的探讨[J].计算机工程与应用,2004,40(1):185.

[13] 孟晓东,袁道华,施惠丰.基于回归模型的数据挖掘研究[J].计算机与现代化,2010(1):26-28.

[14] DELEN D, FULLER C, MCCANN C, et al. Analysis of healthcare coverage: a data mining approach[J]. Expert Systems with Applications, 2009, 36(2):995-1003.

[15] SAMANEH S J, AMTRHASSAN M J, ZAHRA Z J J. A model for adoption of mobile banking services using classification and regression trees[J]. Journal of US-China Public Administration, 2010, 7(11):66-73.