·论著·

基于机器学习构建江西地区缺血性脑卒中风险预测模型

常怀文 姚音

（复旦大学生命科学学院计算生物学系，上海 200438）

【摘要】 目的 借助机器学习构建江西地区缺血性脑卒中风险预测模型。**方法** 借助问卷的方式获取 2020 年 1 月～2020 年 12 月就诊于江西省某三甲医院的 574 例缺血性脑卒中患者及 171 例健康人群调查数据，收集其基本信息及缺血性脑卒中病理特征，使用机器学习分析上述特征关系，并构建缺血性脑卒中风险预测模型。**结果** 基于 t 检验及 *Mann-Whitney* 检验发现，缺血性脑卒中与健康人群年龄、颈动脉狭窄或闭塞是否有症状及收缩压比较差异有统计学意义（$P<0.05$）。同时借助 *AUC* 评估方法，对上述指标基于朴素贝叶斯模型、支持向量机构建江西地区缺血性脑卒中风险预测模型，认为支持向量机表现最优（AUC 分别为 0.996、1.000）。**结论** 本研究所构建的江西地区缺血性脑卒中风险预测模型具有较高的可信度，且缺血性脑卒中与多种病理特征存在较强的相关性，在后续缺血性脑卒中的预防干预及精准医疗过程中应当重点关注。

【关键词】 缺血性脑卒中；朴素贝叶斯；支持向量机

【中图分类号】 R743.34 **【文献标志码】** A **DOI**：10.3969/j.issn.1672-3511.2022.08.017

Constructing risk prediction model of ischemic stroke in Jiangxi based on machine learning

CHANG Huaiwen, YAO Yin

(*Department of Computational Biology, School of Life Sciences, Fudan University, Shanghai* 200438, *China*)

【Abstract】 Objective With the help of machine learning, the risk prediction model of ischemic stroke in Jiangxi is constructed. **Methods** With the help of questionnaires, 574 patients with ischemic stroke and 171 healthy people who visited a third-class hospital in Jiangxi Province from January 2020 to December 2020 were obtained. Their basic information and pathological characteristics of ischemic stroke were collected, and the above characteristics were analyzed by machine learning, and the risk prediction model of ischemic stroke was constructed. **Results** Based on the t-test and Mann Whitney test, it was found that there were significant differences in age, symptoms of carotid stenosis or occlusion, and systolic blood pressure between ischemic stroke and healthy people ($P<0.05$). At the same time, with the aid of AUC evaluation method, based on Naive Bayes model and support vector mechanism, the risk prediction model of ischemic stroke in Jiangxi Province is established for the above indicators, and it is considered that support vector machine performs best (AUC is 0.996 and 1.000 respectively). **Conclusion** The risk prediction model of ischemic stroke in Jiangxi constructed by this study has high reliability, and there is a strong correlation between ischemic stroke and a variety of pathological characteristics, which should be paid attention to in the follow-up prevention and intervention of ischemic stroke and precision medical treatment.

【Key words】 Ischemic Stroke; Naive Bayes; Support Vector Machine

脑卒中包括缺血性脑卒中和出血性缺血性脑卒中两种，其中缺血性脑卒中占 70%～80%，是我国成年人致死和致残的首位原因[1-2]。缺血性脑卒中是由于脑供血动脉（颈动脉和椎动脉）狭窄或闭塞，脑供血不足而引起的脑组织坏死的总称[3]。然而，许多患者无法从早期治疗中获益，大量的时间在院外丢失，这往往是因为缺乏对脑卒中症状的认知，缺乏快速寻求紧急救治的途径，或对其缺乏应急反应[4]。据不完全统计，缺血性脑卒中多发生于中老年人，其主要原因

通信作者：姚音，E-mail：yin_yao@fudan.edu.cn

引用本文：常怀文，姚音. 基于机器学习构建江西地区缺血性脑卒中风险预测模型[J]. 西部医学，2022，34(8)：1182-1186. DOI：10.3969/j.issn.1672-3511.2022.08.017

是长期吸烟、酗酒、肥胖，以及长期控制不佳的高血压、糖尿病、高脂血症，使脑动脉粥样硬化越来越严重[5]。由于病理特征复杂，且影响因素众多，目前临床上难以判定多种因素作用的程度会对缺血性脑卒中的发生产生何种影响，而其前期的预防干预优势又显著大于后期治疗[2-3,5]。如今，随着大数据的深入应用，机器学习已经进入医疗领域，其卓越的算法特性能够更好、更快地帮助我们找出发病源及关联属性，为后期精准医疗提供帮助[6]。因此本研究以选取江西地区的样本数据为例，利用多种机器学习方法构建缺血性脑卒中风险预测模型，以此挖掘缺血性脑卒中的发病机制，为缺血性脑卒中的提前干预与控制提供理论依据。

1 资料与方法

1.1 一般资料 研究使用的数据共 745 例，来自于 2020 年 1 月～2020 年 12 月就诊于江西省某三甲医院的 574 例缺血性脑卒中患者及 171 例健康人群问卷调查数据。纳入标准：①诊断为缺血性脑卒中成年患者。②具备基本的认知能力。③签署知情同意书。其中，非脑卒中数据以问卷形式在江西地区发放，抽取无心脑血管疾病的健康人群问卷作为调查样本；缺血性脑卒中患者均为首次就诊，符合条件的参与者在症状出现后 48 h 内通过 CT 或 MRI 证实为缺血性脑卒中，并且收缩压升高在 140～220 mmHg。该研究获得了医院机构委员会的伦理批准(CKLL-2018005)，所有研究活动均按照其指导方针进行。在详细解释本研究的性质后，从所有研究参与者处获得书面知情同意书。

1.2 方法

1.2.1 指标选取 由于人群特征的不同，不同地区心血管疾病危险因素的选择存在一些差异，但相关专业人士一致认为主要危险因素应符合以下标准[7]：①于许多人群中的存在率很高。②对心血管疾病的风险有重要的独立影响。③经过治疗和控制，可以降低风险。基于此，本研究对用于调查的影响心血管疾病发生的重要生物学指标进行介绍并对其进行数据编码，见表 1。

1.2.1.1 性别和年龄 影响心血管疾病的不可控制的危险因素之中主要包含性别和年龄。一般而言，心血管疾病的风险随着年龄的增长而增加[2-3,6]。研究表明，男性心血管疾病的风险高于女性，同时，随着年龄的增长，心血管疾病复发率的性别差异呈现逐渐减弱的迹象[5,7]。

1.2.1.2 高血压 根据 JNC-VII，年龄在 40～70 岁且血压为(115～185)/(75～115) mmHg 的个体，收缩压(SBP)每增加 20 mmHg 和舒张压(DBP)每增加 10 mmHg，患心血管疾病的危险性将以倍数的形式提升[8]。高血压最终会导致心脏、大脑、肾脏和外周血管发生病理变化，从而导致一系列并发症，例如充血性心力衰竭、左心室肥大、冠心病及脑血管疾病等并发症[9]。

表 1 定性指标的选取及变量定义

Table 1 Selection of qualitative indicators and definition of variables

性别	年龄	颈动脉狭窄或闭塞是否有症状	是否有狭窄	收缩压(mmHg)
				95 及以下 0
	25 岁及以下 0			95～105 1
	25～35 岁 1			105～113 2
	35～45 岁 2			114～118 3
男 0	45～55 岁 3	无 0	无 0	118～124 4
女 1	55～65 岁 4	有 1	有 1	124～130 5
	65～75 岁 5			130～136 6
	75～85 岁 6			136～145 7
	85 岁及以上 7			145～155 8
				155 及以上 9

2 结果

2.1 指标分布差异性分析及可视化 本研究考虑在患缺血性脑卒中与健康两种情况下对比上述所选指标间是否存在显著差异性。对于参数检验，其假定数据可以由一个或多个参数定义的分布很好地描述，且在大多数情况下是通过正态分布来描述的[10]。如果样本数据集无法被选择的分布近似的时候，参数检验的结果会存在极大的误差，此时应当考虑非参数检验[11]。因此，本研究对所选指标绘制概率密度函数图，发现收缩压的分布不具有正态性，呈现偏态分布，见图 1。

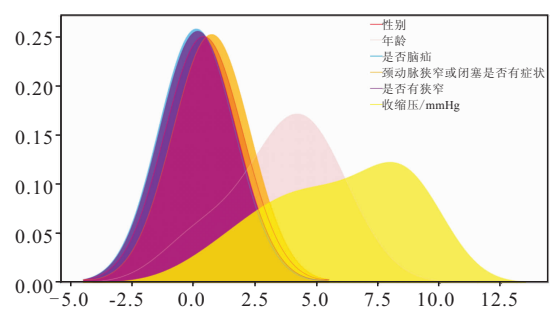

图 1 所选指标的概率密度函数图

Figure 1 Probability density function diagram of selected indicators

2.2 差异性检验

2.2.1 参数检验 本研究对患缺血性脑卒中与健康两种情况下服从正态分布的指标进行独立组别之间的 t 检验。显著性水平为 $\alpha=0.05$，检验统计量为

$$t = \frac{\bar{x} - \bar{y}}{s_w \sqrt{+}} \quad (1)$$

式(1)中 $\bar{x}$ 和 s_x^2 分别为来自患脑卒情况下的正态总体 $N(\mu_1,\sigma_1^2)$ 的样本 $x_1,\cdots,x_m$ 的均值和方差；$\bar{y}$ 和 s_y^2 分别为来自健康情况下的正态总体 $N(\mu_2,\sigma_2^2)$ 的样本 $y_1,\cdots,y_n$ 的均值和方差，其中 $s_w^2=\frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2}$。检验结果表明，患缺血性脑卒中与健康两种情况下年龄及颈动脉狭窄或闭塞是否有症状这两个指标间存在显著差异，检验结果见表 2。

表 2 正态性指标独立组别之间的 *t* 检验结果

Table 2 The t test results between independent groups of normality index

指标	*P-value*
性别	0.503
年龄	1.882×10^{-100}
颈动脉狭窄或闭塞是否有症状	0.000
是否狭窄	0.783

2.2.2 非参数检验 本研究对患缺血性脑卒中与健康两种情况下不服从正态分布的收缩压指标进行 *Mann-Whitney* 检验。显著性水平为 $\alpha=0.05$，检验假设为

$$H_0:F(x)=G(y);H_1:F(x)\neq G(y) \quad (2)$$

H_0 的拒绝域为

$$T=\sum_{i=1}^{n}R_i\leqslant r_1 or T\geqslant r_2 \quad (3)$$

其中 r_1 与 r_2 由下式算出：

$$\begin{cases} r_1\approx\frac{1}{2}m(m+n+1)-z_{0.05}\sqrt{\frac{mn(m+n+1)}{12}} \\ r_2\approx\frac{1}{2}m(m+n+1)+z_{0.05}\sqrt{\frac{mn(m+n+1)}{12}} \end{cases} \quad (4)$$

式(2)中 $F(x)$ 与 $G(y)$ 分别表示患缺血性脑卒中与健康两种情况下指标的分布函数，式(3)中 $\sum_{i=1}^{n}R_i$ 表示将两类数据按照大小混合排列后单类数据的秩和，式(4)中 m 和 n 分别表示患缺血性脑卒中与健康两种情况下的样本量大小。检验结果表明，u 统计量为 14539.5，p 值为 3.106×10^{-56}，也即患缺血性脑卒中与健康两种情况下患者的收缩压指标存在显著差异。上述检验结果表明，在患缺血性脑卒中与健康两种情况下年龄、颈动脉狭窄或闭塞是否有症状及收缩压之间将会存在显著差异，这将对后续缺血性脑卒中的预防干预以及精准医疗提供有力的理论支撑。

2.3 基于朴素贝叶斯的缺血性脑卒中风险预测模型构建

2.3.1 基本模型 朴素贝叶斯法对条件概率分布的条件独立性假设为

$$P(X=x\mid Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\mid Y=c_k) \quad (5)$$

针对所得输入 x，后验概率分布 $P(Y=c_k\mid X=x)$ 可由学习模型计算所得。与此同时，x 的类输出可由后验概率最大的类所得。根据贝叶斯定理计算后验概率：

$$P(Y=c_k\mid X=x)=\frac{P(X=x\mid Y=c_k)P(Y=c_k)}{\sum_k P(X=x\mid Y=c_k)P(Y=c_k)} \quad (6)$$

式(5)代入式(6)，有

$$P(Y=c_k\mid X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)} \quad (7)$$

也即朴素贝叶斯分类公式。

2.3.2 模型构建 基于朴素贝叶斯分类器的高效性，其通过单独查看每个特征来学习参数，并从每个特征中收集简单的类别统计数据，且本研究所选指标数据大多为二分类数据，考虑使用 *BernoulliNB* 进行机器学习，其假设缺血性脑卒中特征的先验概率为多项式分布，即为

$$P(X_j=x_{jl}\mid Y=c_k)=\frac{x_{jl}+\lambda}{m_k+n\lambda} \quad (8)$$

其中，$P(X_j=x_{jl}\mid Y=c_k)$ 为第 k 个类别的第 j 维特征的取值条件概率，m_k 为训练集中第 k 类输出的样本数量，λ 为大于 0 的常数，通常等于 1，即拉普拉斯平滑[12]。

由于 BernoulliNB 含有一个参数 α(即上述拉普拉斯平滑参数 λ)，用于控制模型复杂度。α 的工作原理是，算法向数据中添加 α 这么多的虚拟数据点，这些点对所有特征都取正值。这可以将统计数据“平滑化”。α 越大，平滑化越强，模型复杂度就越低。另一方面，算法的性能对 α 值的鲁棒性相对较好。需要强调的是，调整 α 将会使得精度略有提高。本研究分别选取 $\alpha=1$，$\alpha=10$，$\alpha=100$ 构建伯努利朴素贝叶斯分类器，其中模型精度见表 3。

表 3 对于不同的 *α* 值的伯努利朴素贝叶斯分类器在缺血性脑卒中数据集上的模型精度

Table 3 For different substances α Model accuracy of Bernoulli naive Bayes classifier with variable value on ischemic stroke data set

α	训练集精度	测试集精度
1	0.8351	0.8226
10	0.8351	0.8172
100	0.7670	0.7581

由于 $\alpha=1$ 时模型精度最高，故本研究最终选取考

虑先验概率且 $\alpha=1$ 的伯努利朴素贝叶斯分类器。

2.4 基于支持向量机的的缺血性脑卒中风险预测模型构建

2.4.1 基本模型 本文给出线性可分支持向量机学习算法步骤：输入：线性可分训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in X=R^n$，$y_i\in Y=\{-1,+1\}$，$i=1,2,\cdots,N$。

输出：分类决策函数以及分离超平面。

(1)构造并求解约束最优化问题

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$
$$s.t.\sum_{i=1}^{N}\alpha_i y_i=0\alpha_i\geqslant 0,i=1,2,\cdots,N \quad (9)$$

求得最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^T$，其中 $\alpha_i\geqslant 0$ 为对偶算法中的拉格朗日乘子。

(2)计算

$$\omega^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i \quad (10)$$

并选择 α^* 的一个正分量 $\alpha_j^*>0$，计算

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j) \quad (11)$$

(3)求得分离超平面

$$w^*\cdot x+b^*=0 \quad (12)$$

分类决策函数：

$$f(x)=sign(w^*\cdot x+b^*) \quad (13)$$

2.4.2 模型构建 对于所得样本数据集，由于样本在不同特征上的维度不同，将会使得距离计算存在问题，则考虑先进行标准化[13]。同时考虑其具有较强的线性可分性，故使用线性支持向量机。要注意的是，对于线性支持向量机，与 *Logistic* 回归类似，具有决定正则化强度的权衡参数 C[11,14]。本研究分别选取 C 值 0.001、1.000 和 100.000 构建线性支持向量机，模型精度见表 4。

表 4 对于不同的 C 值，线性支持向量机在缺血性脑卒中数据集上的模型精度

Table 4 Model accuracy of linear support vector machine on ischemic stroke data set for different exergy C exergy values

C	训练集精度	测试集精度
0.001	0.9839	0.9839
1.000	0.9910	0.9892
100.000	0.9910	0.9839

由于 $C=1.000$ 时模型精度最高，故本研究最终选取 $C=1.000$ 的线性软间隔支持向量机，其分类决策函数见式(14)，其中 x_j，$j=1,\cdots,5$ 为缺血性脑卒中数据集输入变量。

$$f(x)=sign(0.06x_1+0.54x_2+1.42x_3-0.14x_4+0.63x_5+1.60) \quad (14)$$

2.5 基于 *ROC* 与 *AUC* 的模型选择 *ROC* 曲线考虑给定分类器的所有可能阈值，并显示假正例率和真正例率，而不是报告准确率和召回率[15-16]。对于 *ROC* 曲线，理想曲线应接近左上角。本研究希望分类器的召回率很高，同时使假正例率很低。利用 AUC 分数来比较朴素贝叶斯模型及支持向量机，本研究发现支持向量机的表现比朴素贝叶斯模型要略好一些，见表 5。综上所述，本研究最终选取支持向量机模型(C=1.000)作为缺血性脑卒中风险预测模型。

表 5 两种自动化模型的 AUC

Table 5 AUC of the two automation models

模型	AUC
朴素贝叶斯模型	0.996
支持向量机	1.000

3 讨论

3.1 缺血性脑卒中的现状分析 本研究的研究重点在于缺血性脑卒中前期风险预测及干预。基于 745 例样本数据，借助 *t* 检验及 *Mann-Whitney* 检验发现，在患缺血性脑卒中与健康两种情况下年龄($P=0.000$)、颈动脉狭窄或闭塞是否有症状($P=0.000$)及收缩压($u=14539.500$，$P=0.000$)间存在显著差异，在后续缺血性脑卒中的预防干预以及精准医疗过程中应当重点关注。

3.2 缺血性脑卒中风险预测模型 缺血性脑卒中具有发病率高、致残率高、死亡率高和复发率高的特点。因此，在源头控制缺血性脑卒中发病率及风险干预尤为重要[17]。因此，本研究对所得样本数据进行全方位、多角度的挖掘分析，利用机器学习方法构建风险预测模型并进行严格的模型选择力求获得最优的缺血性脑卒中风险预测模型。

在缺血性脑卒中的预测研究中，许多已构建的稳定的评分方法与预测模型普遍建议选择改良的弗明汉缺血性脑卒中量表、汇集队列方程、缺血性脑卒中风险计算器等工具进行缺血性脑卒中风险评估，但是这些模型主要针对欧美人群，对我国人群的缺血性脑卒中风险评估预测效果不佳[18-20]。同时需要注意的是，上述模型虽然容易被理解，但准确性不高、误差较大。本研究以江西地区的案例为基础，从中国缺血性脑卒中实际病理情况出发，通过数据挖掘的方式揭示影响缺血性脑卒中的危险因素，同时探讨各因素之间的分布特征及相关性，并选取少数典型指标用于后续建模，这极大降低了计算复杂度与算法迭代次数，同

时使得模型更准确、误差小且区分度高，更容易被理解[19,21-23]。

另一方面，凭借机器学习的卓越特性，本研究还利用 ROC 与 AUC 进行模型筛选。需要强调的是，由于本研究所得到的为不平衡数据，认为 AUC 是一个比精度好得多的指标。AUC 等价于从正类样本（患有缺血性脑卒中）中随机挑选一个点，由分类器给出的分数比从反类样本（健康）中随机挑选一个点的分数更高的概率[21-24]；可以被解释为评估正例样本的排名[25]。本研究所训练支持向量机模型的 AUC 为 1.000，说明所有正类点的分数高于所有反类点。基于此，本研究认为对于所得不平衡的缺血性脑卒中数据集，使用 AUC 进行模型选择比使用精度更有意义。本研究最终从数理方向论证支持向量机为最优缺血性脑卒中风险预测模型，所构建模型具有较高的准确性，在一定程度上极大地确保了本研究的完备性及可靠性，提高了最初无症状人群中缺血性脑卒中预测干预的准确性。

4 结论

本研究所构建的基于现阶段的符合我国国情的缺血性脑卒中风险预测模型可在源头上控制缺血性脑卒中发病率，并通过采取及时且准确的干预措施，极大程度地确保国民的健康。

【参考文献】

[1] 刘泽文. 基于机器学习的缺血性脑卒中复发预测模型研究[D]. 长沙：湖南大学，2015：1-35.

[2] 项钰. 南方人群心血管疾病患者危险因素的调查研究[D]. 广州：华南理工大学，2011：10-20.

[3] 王陇德，刘建民，杨弋，彭斌，王伊龙. 我国缺血性脑卒中防治仍面临巨大挑战[J]. 中国循环杂志，2019，2(5)：102-104.

[4] 侯玉梅，曾慧等. 基于数据挖掘的缺血性脑卒中患病风险预测[J]. 中国老年学杂志，2021，1(5)：35-38.

[5] 巢宝华，刘建民，王伊龙，杨弋，彭斌等. 中国缺血性脑卒中防治：成就、挑战和应对[J]. 中国循环杂志，2019，7(11)：45-48.

[6] 王晓峰，孙碧竹. 农村留守老人健康管理模式构建[J]. 社会科学战线，2019，4(5)：103-108.

[7] 汤少梁，沈旖旎. 健康扶贫视角下基于客户关系管理的慢性病管理体系研究[J]. 中国全科医学，2018，3(32)：13-15.

[8] 王陇德，毛群安，张宗久，等. 我国缺血性脑卒中防治仍而临巨大挑战——《中国缺血性脑卒中防治报告 2018》概要[J]. 中国循环杂志，2019，34(2)：105-119.

[9] 王禹婷，雷亚莉，祝愿，等. 成都市某医院体检人群缺血性脑卒中相关知识的调查研究[J]. 成都医学院学报，，2018，13(3)：279-283.

[10] 魏芳，易小萍，桂金艳，等. 成都地区缺血性脑卒中住院患者对卒中症状及急救知识的认知[J]. 中国实用神经疾病杂志，2015，18(10)：57-58.

[11] CARROLL C，HOBART J，FOX C，*et al*. Stroke in Devon：knowledge was good，but action was poor[J]. Journal of Neurology and Psychiatry，2004，75(4)：567-571.

[12] GREENLUND K J，NEFF L J，ZHENG Z J，*et al*. Low public recognition of major stroke symptoms[J]. Am J Prev Med，2003，25(4)：315-319.

[13] Montaner J，Vidal C，Molina C，*et al*. Selecting the target and the message for a stroke public education campaign：a local survey conducted by neurologists[J]. Eur J Epidemiol，2001，17(6)：581-586.

[14] 吕力兢. 基于卷积神经网络的结肠病理图像中的腺体分割[D]. 南京：东南大学，2016：23-26.

[15] 刘铁钧. 基于神经网络算法的电网短期负荷预测在兴隆电力公司的应用研究[D]. 北京：华北电力大学，2017：13-18.

[16] 徐谦. 基于半监督方法的生物医学事件抽取的研究[D]. 大连：大连理工大学，2013：19-30.

[17] 朱艳琳. 基于文本挖掘的多准则推荐系统[D]. 天津：天津大学. 2016：11-32.

[18] 贾博轩. 基于手机传感器的人类复杂性为识别方法的研究[D]. 黑龙江大学. 2015：9-17.

[19] 张兴平. 基于 Hadoop 的微博用户感情分类研究与实现[D]. 西安：西安电子科技大学，. 2014：19-32.

[20] 玄英花. 医学统计学教学中利用 R 语言进行描述性统计分析[J]. 教育教学论坛，2016，6(2)：105-119.

[21] 余健浩. 基于支持向量回归集成的蛋白质-ATP 绑定点预测研究[D]. 南京：南京理工大学，2015：18-34.

[22] 殷敏，李晓辉，李常宝，顾平莉，张可，吕守业. 一种法律判决预测的影响因素分析方法[J]. 计算机与现代化，2021，4(2)：11-19.

[23] 王怀亮. 箱须图在识别统计数据异常值中的作用及 R 语言实现[J]. 商业经济，2011，3(2)：25-29.

[24] 赵磊. 基于 Pawlak 属性重要度的混合感情特征选择算法研究[D]. 昆明：云南财经大学，2014：29-39.

[25] 陆晓，徐鹏，冯树海，等. 考虑多时段数据随机误差的变压器正序参数辨别与评估方法[J]. 电网技术，2019，3(2)：123-134.

（**收稿日期：2021-12-22；修回日期：2022-06-04；编辑：黎仕娟**）