

基于时间序列模糊分割的通信数据 分类算法设计

张勇飞¹, 陈 涛²

(1. 南昌大学 科学技术学院, 江西 九江 332020; 2. 南昌大学 数学与计算机学院, 江西 南昌 330031)

摘 要:为解决通信数据集不平衡性明显导致分类难度大幅上升的问题,提出一种时间序列模糊分割的分类算法。利用主成分分析法取得特征值最大的特征矢量,根据区间数理论建立数据的时间序列,用数据与类别间的兰氏距离测度函数表征通信数据时间序列的分割目标,依据模糊分类矩阵差值与收敛条件的判定关系,获取模糊分割的分类结果。实验结果表明:本文算法 3 种分类情况的数据数量始终位于相应的理想水平上,正确性较高,错误性与失效性较低。

关键词:时间序列;模糊分割;通信数据;分类算法;主成分分析法;区间模糊数

中图分类号:TP391 **文献标志码:**A **文章编号:**1671-5497(2023)11-3268-06

DOI:10.13229/j.cnki.jdxbgxb.20220736

Design of communication data classification algorithm based on fuzzy segmentation of time series

ZHANG Yong-fei¹, CHEN Tao²

(1. College of Science and Technology, Nanchang University, Jiujiang 332020, China; 2. School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, China)

Abstract: Due to the obvious unbalance of communication data set, the difficulty of classification is greatly increased. Therefore, a time series fuzzy segmentation classification algorithm is proposed. Using the principal component analysis method, the feature vector with the largest eigenvalue is obtained, the time series of the data is established according to the interval number theory, and the segmentation target of the time series of the communication data is characterized by the Langley distance measure function between the data and the class. The classification result of the fuzzy segmentation is obtained according to the judgment relationship between the difference of the fuzzy classification matrix and the convergence condition. The experimental results verify that the data quantity of the three classification cases is always at the corresponding ideal level, with high accuracy and low error and failure.

Key words: time series; fuzzy segmentation; communication data; classification algorithm; principal component analysis; interval fuzzy number

收稿日期: 2022-06-13.

基金项目: 国家自然科学基金项目(71363043); 2021 年度江西省教育厅科学技术研究项目(GJJ217813).

作者简介: 张勇飞(1981-), 男, 副教授, 博士. 研究方向: 数学与应用数学. E-mail: zyf3201002@163.com

通信作者: 陈涛(1963-), 男, 教授, 博士. 研究方向: 概率论与数理统计, 应用统计及数学建模. E-mail: zyf5353@163.com

0 引言

国民经济飞速发展,移动互联网^[1]、物联网^[2]日益普及。随着数据采集手段与渠道不断增加,数据规模呈爆炸式增长,其中以通信数据^[3]最为庞大。当各行各业迈入数字化与信息化的进程后,数据价值与日俱增,在各大重要领域中享有较高的地位与关注度,是企业运用数据挖掘技术^[4]来判定现状、制定决策的重要依据。作为数据挖掘技术的关键处理环节,数据分类结果直接影响挖掘的精准性。当前处理的大部分数据都存在显著的不平衡特征^[5],该特征将因少数类样本数据过少而加剧偏差的发生概率,增加数据分类难度,因此,如何分类具有不平衡特征的数据集逐渐成为相关领域的热点研究课题。

近几年较为主流的不平衡数据集分类方法主要有:侯贝贝等^[6]提出基于边界混合重采样的非平衡数据分类方法,该方法引入支持 k -离群度概念,找出数据集中的边界点集和非边界点集,利用改进的SMOTE算法将少数类中的边界点作为目标样本合成新的点集,同时对多数类中的非边界点采用基于距离的欠采样算法,以此达到类之间的平衡;李锦珑等^[7]提出基于改进果蝇优化算法的不平衡数据分类方法,并利用此果蝇优化算法良好的全局和局部搜索性能,对类别不平衡数据中样本的误分类代价进行了优化。尽管以上方法均取得了良好成效,但对于特征值持续发生变化的动态通信数据并不具有较好的分类能力。因此,本文将通信数据转换成时间序列,设计出一种模糊分割的分类算法,并尝试通过以下手段,进一步提升分类精准性:利用主成分分析法,降维处理通信数据;正交分解协方差矩阵,去除数据特征间的关联性;基于区间数理论,建立数据的时间序列;以Gath-Geva为基本算法,用数据与类别的间距函数表征通信数据时间序列的分割目标。进行了分类实验,经实验验证可知,本文方法3种分类情况的数据数量始终位于相应的理想水平上,正确性较高,错误性与失效性较低。所设计的时间序列模糊分割分类算法能有效降低数据的误分概率,分类效果较好,再次验证了本文方法对于通信数据的可行性与实践性。

1 通信数据预处理

利用移动终端^[8]采集通信数据,初始数据集

中包含数据上传时间、连接基站、信号强度、热点频率等多种属性的数据信息。为加快分类速度、提高分类精准度,采用无监督的主成分分析法^[9],将数据映射至取得的数据分布最优子空间中,完成通信数据的降维处理。

假设由 n 个数据特征 f 组成的通信数据集是 $F = \{f_1, f_2, \dots, f_n\}$,第 j 个特征 f_j 与第 $j+1$ 个特征 f_{j+1} 的期望值分别是 $E(f_j)$ 、 $E(f_{j+1})$,则两特征间的协方差计算公式如下:

$$\text{cov}(f_j, f_{j+1}) = E \left[\frac{f_j - E(f_j)}{f_{j+1} - E(f_{j+1})} \right] \quad (1)$$

由此推导出数据集 F 的属性协方差矩阵表达式,如下所示:

$$\text{cov}(F) = \begin{Bmatrix} \text{cov}(f_1, f_1) & \cdots & \text{cov}(f_n, f_1) \\ \vdots & & \vdots \\ \text{cov}(f_1, f_n) & \cdots & \text{cov}(f_n, f_n) \end{Bmatrix} \quad (2)$$

正交分解上式协方差矩阵 $\text{cov}(F)$,去除数据特征间的关联性,加强降维处理效果。所得矩阵分解式如下所示:

$$\text{cov}(F) = Q^T \otimes Q \quad (3)$$

式中: Q 为正交矩阵,由协方差矩阵 $\text{cov}(F)$ 中各特征矢量构成; Q^T 表示转置的正交矩阵。

降序排列各矢量对应的特征值,得到前几个拥有最大数值的特征矢量;根据所得矢量明确数据分布的最优子空间,在其中完成初始数据的投影操作,获取经过降维处理的新数据。

通信数据特征值呈持续变化状态,故基于区间数理论^[10]建立数据的时间序列,以更好地抑制通信数据中模糊的不确定性因素。令模糊的数据属性与变化状态之间呈正比例函数关系,利用绝对变化率描述通信数据的模糊属性。

假定时间 t 对应的通信数据是 z_t ,则通过下列表达式计算出该数据在 t 时的绝对变化率 e_{z_t} :

$$e_{z_t} = \left| \frac{(z_t - z_{t-1})(z_{t-1} - z_{t-2})}{z_t} \right| \quad (4)$$

若 t 时数据 z_t 的一阶差分为 d_{z_t} ,则该数据的模糊半径 L_{z_t} 由下式解得:

$$L_{z_t} = \frac{\sum_{t=1}^{T'} e_{z_t}}{e_{z_t} * |d_{z_t}|} \quad (5)$$

式中: T' 为时序总数;*表示复共轭。

由此推导出用于描述数据 z_i 属性的区间直觉模糊数^[11], 如下所示:

$$B_{z_i} = [z_i - L_{z_i}, z_i + L_{z_i}] \quad (6)$$

同理, 得到所有数据的区间直觉模糊数, 建立由各区间直觉模糊数组成的时间序列 $F(t) = \{B_{z_0}, B_{z_1}, \dots, B_{z_i}, \dots, B_{z_{t'}}\}$, 降低后续分割分类难度与复杂度。

2 用于通信数据的时间序列模糊分割分类算法

结合 Gath-Geva 算法^[12], 面向构建的通信数据时间序列, 设计出时间序列模糊分割分类算法。用数据与类别的兰氏间距测度函数^[13]表征通信数据时间序列的分割目标, 按照如下所述的算法操作流程, 实现数据的时间序列模糊分割分类:

(1) 将分类个数 c 、迭代次数 l 、收敛条件 ϵ 、模糊分类矩阵 U 、最小分割数量 $c_{\min}$ 等参数恢复至初始值。

(2) 已知目标数据集中的第 i ($i \in c$) 个类别是 η_i , 含有 5 个参数, 分别是类别先验概率 $p(\eta_i)$ 、时间序列的簇中心 v_i^t 、数据集的簇中心 v_i^x 、时间序列高斯函数的方差 $\sigma_{i,t}^2$ 、数据集高斯函数的方差 $\sigma_{i,x}^2$, 各参数的计算公式分别如下所示:

$$p(\eta_i) = \sum_{k=1}^n \mu_{i,k} \quad (7)$$

$$v_i^x = \frac{\sum_{k=1}^n \mu_{i,k} * W_i(y_{i,k})}{\sum_{k=1}^n \mu_{i,k}} \quad (8)$$

$$v_i^t = \frac{\sum_{k=1}^n \mu_{i,k} * v_i^x}{\sum_{k=1}^n \mu_{i,k}} \quad (9)$$

$$\sigma_{i,t}^2 = \frac{\sum_{k=1}^n \mu_{i,k} * v_i^t}{\sum_{k=1}^n \mu_{i,k}} \quad (10)$$

$$\sigma_{i,x}^2 = \frac{W_i}{M_i} \quad (11)$$

式中: $y_{i,k}$ 为隐变量^[14]; M_i 为类别矩阵, 更新值为 W_i ; $\mu_{i,k}$ 为第 k 个数据分类至类别 η_i 的隶属度^[15]。

(3) 假设数据集中第 k 个数据是 z_k , 组成部分为两个相互独立的变量, 分别是时间序列的时间

点 t_k 与数据集的数据点 x_k , 数据 z_k 与类别 η_i 的兰氏间距测度是 $D^2(z_k, \eta_i)$, 被划分到该类别 η_i 的概率是 $p(z_k | \eta_i)$ 。将概率知识^[16]作为模糊分割分类算法的理论支撑, 根据测度 $D^2(z_k, \eta_i)$ 与概率 $p(z_k | \eta_i)$ 之间的负相关性, 构建出数据 z_k 与类别 η_i 的兰氏间距测度 $D^2(z_k, \eta_i)$ 计算公式^[17], 如下所示:

$$D^2(z_k, \eta_i) = \frac{p(\eta_i)}{p(z_k | \eta_i)} = \frac{p(\eta_i)}{p(t_k | \eta_i) * p(x_k | \eta_i)} \quad (12)$$

式中: $p(t_k | \eta_i)$ 、 $p(x_k | \eta_i)$ 分别为将时间点 t_k 与数据点 x_k 分别划分到类别 η_i 的概率^[18]。

(4) 利用下列表达式求解出模糊分类矩阵 U :

$$U = [\mu_{i,k}]_{c \times n} = \sum_{a=1}^c \left(\frac{D^2(z_k, \eta_i)}{D^2(z_k, \eta_a)} \right) \quad (13)$$

式中: η_a 为第 a 个类别^[19], $a \neq i$ 且 $a \in c$ 。

若最后两次迭代所得的模糊分类矩阵差值与收敛条件 ϵ 之间呈小于关系, 数据分类任务完成, 得到分割分类结果^[20]; 反之, 则返回第 (2) 步重新求取分类参数, 进入新一次的迭代循环运算。

3 通信数据分类实验

3.1 实验数据集与评估指标选取

从某通信行业近一年的数据记录中, 随机选取 10 个不同规模的数据集作为实验对象, 设定各数据集中数据量较大的类别为正类, 较小的类别为负类。所选数据集包含的样本数据总数量及正类、负类各自含有的样本数据数量, 如表 1 所示。

选取不平衡数据分类研究中常用的 $F1$ 值与

表 1 通信数据集规模信息

Table 1 Scale information of communication datasets 个

数据集	正类	负类	总数量
1	1040	124	1164
2	1185	797	1982
3	963	398	1361
4	353	244	597
5	830	212	1042
6	925	341	1266
7	901	180	1081
8	1113	144	1257
9	819	685	1504
10	700	137	837

G 均值两指标作为实验的综合性评级标准,相应的计算公式分别如下所示:

$$F1 = \frac{2 * Pre * rec}{Pre + rec} \quad (14)$$

$$G-mean = \sqrt{rec * TN_R} \quad (15)$$

式中: Pre 为查准率; rec 为召回率; TN_R 为真负率。

两个评估指标的取值范围均为0到1,数值越大,方法的分类能力越强,分类任务完成得越好,分类效果越精准。

实验验证环节中,利用文献[6]提出的基于边界混合重采样、文献[7]提出的改进果蝇优化算法及本文3种针对不平衡数据集的数据分类方法,分别处理所选的通信数据集,通过对比分析各方法的优势与劣势,以便为今后的工作重点与攻克方向提供可靠的数据基础。

3.2 通信数据分类结果分析

各方法在划分通信数据类别的过程中共产生3种分类情况:正确分类、错误分类、未分类。每种分类情况中,不同方法所产生的数据数量如图1所示,可以看出,对于正、负两类相对平衡但样本数据总数量差异较大的2号数据集与4号数据集,3种方法对小规模的4号数据集的分类结果更加理想;对于样本数据总数量差异较小但正、负两类不平衡程度较大的8号数据集与9号数据集,3种方法对不平衡程度较小的9号数据集的分类结果更加理想;无论规模大小,不平衡程度如何,本文方法3种分类情况的数据数量始终位于相应的理想水平上,正确性较高,错误性与失效性较低。

所得实验结论说明,数据集所含样本数据总数量越多,正类与负类样本数据数量差异越大,各方法的正确分类数据数量越少,错误分类与未分类的数据数量越多,两者对分类结果具有较大的干扰性,但本文根据区间数理理论,利用绝对变化率描述通信数据的模糊属性,建立出由各区间直觉模糊数组成的时间序列,更好地抑制了通信数据中模糊的不确定性因素,使数据规模与类别不平衡程度所造成的负面影响大幅下降,故表现出显著的分类优越性。

为更好地体现方法对不平衡数据集的分类效果,提升实验数据的意义与价值,采用 $F1$ 值与 G 均值两项指标,进一步客观评估分类结果的精准度。3种方法的指标数值对比情况如图2所示,可

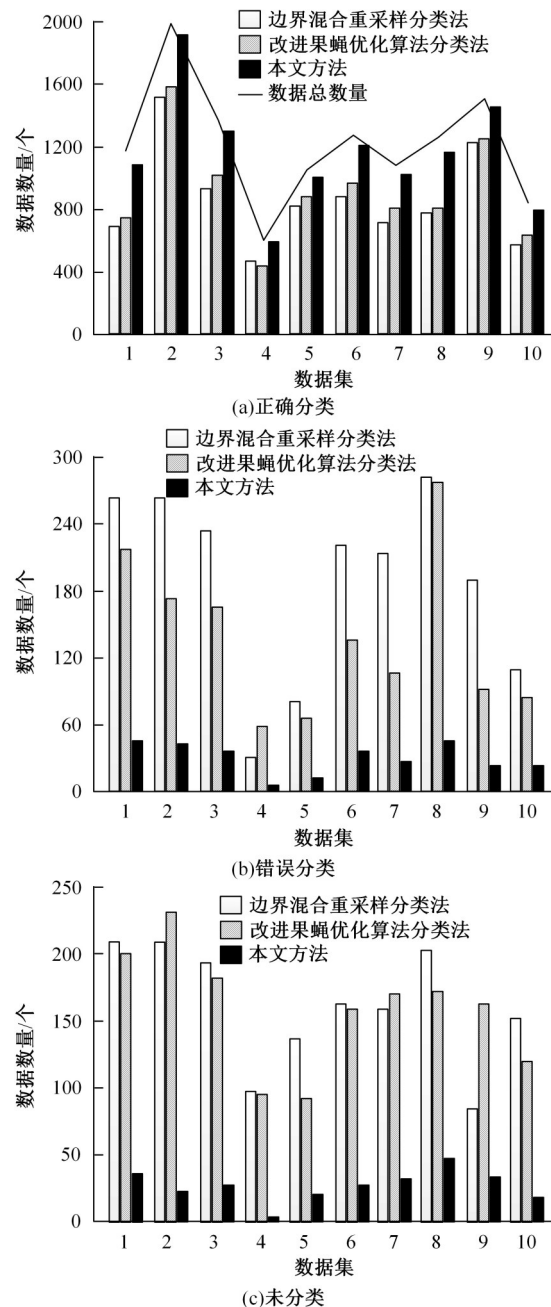


图1 不同方法的通信数据分类结果示意图

Fig. 1 Schematic diagram of the communication data classification results of different methods

以看出,数据集规模越大,正类与负类之间存在不平衡性越大,各方法的评估指标数值均越小。本文方法的指标评估结果远优于对比方法,且始终在较高的水平范围内平稳波动。

该实验环节得到的结论足以说明,所设计的时间序列模糊分割分类算法能有效降低数据的误分概率,分类效果较好,再次验证了本文方法对于通信数据的可行性与实践性。

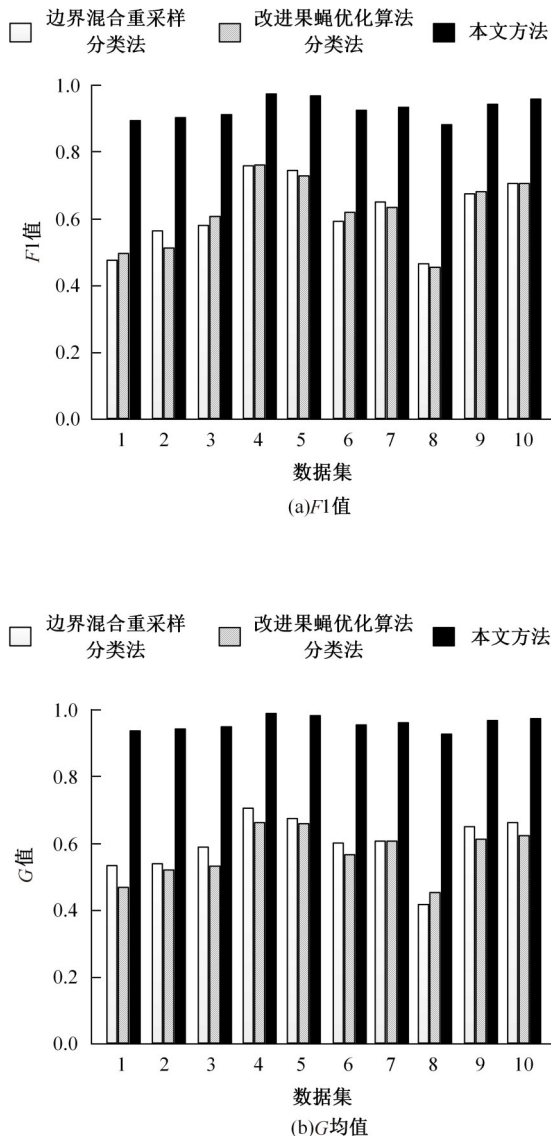


图2 不同方法的数据分类效果评估示意图

Fig. 2 Schematic diagram of data classification effect evaluation of different methods

4 结束语

数据分类是实现挖掘技术的有效手段之一,但当数据因不可控因素造成分布不均时,就会增加分类难度。通信数据即分布不平衡的数据集,为此,本文构建一种时间序列模糊分割算法,用于分类不平衡的通信数据集,本文方法的指标评估结果远优于对比方法,且始终在较高水平范围内平稳波动。尽管本文已经取得一定成果,但仍存在要完善的地方,今后的工作重点主要包含以下几个方面:①针对数据的同构与异构形式,探讨分类算法的处理效果,加强分类算法对数据结构的

免疫性;②探索算法对于其他类型不平衡数据的分类能力,拓宽算法的应用范围与前景。

参考文献:

- [1] 黄燧,刘珍,王若愚,等.移动互联网流量分类的多特征集和多类别标签研究[J]. 计算机应用研究, 2020, 37(11): 3353-3358.
Huang Yi, Liu Zhen, Wang Ruo-yu, et al. Research on multi-feature sets and multi-class labels for mobile Internet traffic classification[J]. Computer Application Research, 2020, 37(11): 3353-3358.
- [2] Huang W, Zhou S, Zhu T, et al. Privately publishing Internet of things data: bring personalized sampling into differentially private mechanisms[J]. IEEE Internet of Things Journal, 2021, 9(1): 80-91.
- [3] Li L, Zhang X, Yue W, et al. Cooperative search for dynamic targets by multiple UAVs with communication data losses[J]. ISA Transactions, 2021, 114(4): 230-241.
- [4] Mai K, Saeed W, Medi I. Churn prediction in digital game-based learning using data mining techniques: logistic regression, decision tree, and random forest[J]. Applied Soft Computing, 2022, 118(1): 108491.
- [5] Shen F, Liu Y, Wang R, et al. A dynamic financial distress forecast model with multiple forecast results under unbalanced data environment[J]. Knowledge-Based Systems, 2020, 192: 105365.
- [6] 侯贝贝,刘三阳,普事业.基于边界混合重采样的非平衡数据分类方法[J]. 计算机工程与应用, 2020, 56(1): 46-52.
Hou Bei-bei, Liu San-yang, Pu Shi-ye. Imbalanced data classification method based on boundary mixed resampling[J]. Computer Engineering and Applications, 2020, 56(1): 46-52.
- [7] 李锦珑,包理群,周彬.改进果蝇优化算法及其在不平衡数据分类中的应用[J]. 西北师范大学学报: 自然科学版, 2021, 57(3): 57-61.
Li Jin-long, Bao Li-qun, Zhou Bin. Improved fruit fly optimization algorithm and its application in imbalanced data classification[J]. Journal of Northwest Normal University (Natural Science Edition), 2021, 57(3): 57-61.
- [8] 陈清,张程.区间约束下移动终端用户偏好的数据挖掘模型[J]. 计算机仿真, 2020, 37(11): 320-323.
Chen Qing, Zhang Cheng. Data mining model of mobile terminal user preference under interval constraints[J]. Computer Simulation, 2020, 37(11): 320-323.
- [9] Huang L, Zhang Q, Zhang L, et al. Efficiency enhancement for underwater adaptive modulation and

- coding systems: via sparse principal component analysis[J]. IEEE Communications Letters, 2020, 24(8): 1808-1811.
- [10] 张海燕, 武晓蓓, 李娜. 基于非对称分布散点分位数的区间数据回归分析[J]. 统计与决策, 2022, 38(2): 26-32.
- Zhang Hai-yan, Wu Xiao-bei, Li Na. Regression analysis of interval data based on scattered quantiles of asymmetric distribution[J]. Statistics and Decision-Making, 2022, 38(2): 26-32.
- [11] 乔剑敏, 李沃源, 赵新平, 等. 一种有方案偏好的 TOPSIS 区间直觉模糊多属性决策方法[J]. 数学的实践与认识, 2020, 50(6): 322-328.
- Qiao Jian-min, Li Wo-yuan, Zhao Xin-ping, et al. A TOPSIS interval intuitionistic fuzzy multi-attribute decision-making method with scheme preference[J]. Practice and Understanding of Mathematics, 2020, 50(6): 322-328.
- [12] Wu X, Zhou H, Wu B, et al. A possibilistic fuzzy gath-geva clustering algorithm using the exponential distance[J]. Expert Systems with Applications, 2021, 184(5): 115550.
- [13] 沙秀艳, 尹传存, 徐泽水. 基于指数熵加权的降维犹豫模糊兰氏距离测度及应用[J]. 控制与决策, 2020, 35(3): 728-734.
- Sha Xiu-yan, Yin Chuan-cun, Xu Ze-shui. Dimensionality reduction hesitant fuzzy rankine distance measurement and application based on exponential entropy weighting[J]. Control and Decision, 2020, 35(3): 728-734.
- [14] 何新林, 戚宗锋, 李建勋. 基于隐变量后验生成对抗网络的不平衡学习[J]. 上海交通大学学报, 2021, 55(5): 557-565.
- He Xin-lin, Qi Zong-feng, Li Jian-xun. Imbalanced learning based on latent variable posterior generative adversarial networks[J]. Journal of Shanghai Jiaotong University, 2021, 55(5): 557-565.
- [15] Akram M, Ullah I, Alharbi M G. Methods for solving LR-Type pythagorean fuzzy linear programming problems with mixed constraints[J]. Mathematical Problems in Engineering, 2021(1): 1-29.
- [16] Lochbihler A. A mechanized proof of the max-flow min-cut theorem for countable networks with applications to probability theory[J]. Journal of Automated Reasoning, 2022(1): 1-26.
- [17] Dakulagi V, Bakhar M. Advances in smart antenna systems for wireless communication[J]. Wireless Personal Communications, 2020, 110(2): 931-957.
- [18] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation[J]. BMC Genomics, 2020, 21(6): 1-13.
- [19] 翁渊瀚, 李南. 基于时间序列模型的文本数据压缩存储算法[J]. 吉林大学学报: 工学版, 2023, 53(7): 2109-2114.
- Weng Yuan-han, Li Nan. Text data compression storage algorithm based on time series model[J]. Journal of Jilin University(Engineering and Technology Edition), 2023, 53(7): 2109-2114.
- [20] 宋震, 柳杰. 旋转机械振动频率时间序列预测算法[J]. 吉林大学学报: 工学版, 2022, 52(8): 1764-1769.
- Song Zhen, Liu Jie. Vibration frequency time series prediction algorithm for rotating machinery[J]. Journal of Jilin University(Engineering and Technology Edition), 2022, 52(8): 1764-1769.