ISGS:一种面向滞后效应的组合模型研究

冯婷婷,彭 岩,王 洁
(首都师范大学管理学院,北京 100056)

摘 要: 针对滞后效应明显、样本量小的数据集,为解决单一算法模型预测精度低、泛化能力差的问题,提出了一种基于等距特征映射算法(Isometric Feature Mapping, ISOMAP)、少数类过采样技术(Synthetic Minority Oversampling Technique, SMOTE)、遗传算法(Genetic Algorithm, GA)、支持向量回归(Support Vector Regression, SVR)的组合模型ISGS(ISOMPA-SMOTE-GA-SVR). 首先,利用ISOMAP和SMOTE算法对滞后效应明显、样本量较小的数据集进行特征变换. 其次,利用SVR算法较强的非线性分类能力及泛化能力对数据集进行回归分析. 最后,利用GA算法对SVR算法的参数进行优化,以提升模型的预测精度. 采用气象因素、空气质量、呼吸系统发病人数三组数据集,基于ISGS模型进行了发病人数预测的仿真实验和对比实验. 实验结果表明,该模型预测精度和准确率较传统模型均有所提高,预测精度达到93.65%(传统单一模型83.481%). 同时具有更好的泛化能力,能够更好地处理高维度、小样本数据集.

关键词: 等距特征映射算法;少数类过采样技术;遗传算法;支持向量回归;组合模型
基金项目: 全国教育科学规划-教育部重点课题(No.DLA190426)
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2023)09-2504-06
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220238

ISGS: A Combinatorial Model for Hysteresis Effects

FENG Ting-ting, PENG Yan, WANG Jie
(*School of Management*, *Capital Normal University*, *Beijing* 100056, *China*)

Abstract: In anticipation of data sets with small sample size and evident lag effects, a novel ISGS (ISOMPA-SMOTE-GA-SVR) model was proposed to address the issues of low prediction accuracy and inadequate generalization in single-algorithm prediction models. This ISGS model integrates isometric feature mapping (ISOMAP), synthetic minority oversampling technique (SMOTE), genetic algorithm (GA), and support vector regression (SVR), thereby providing a comprehensive solution. Firstly, ISOMAP and SMOTE were employed to perform feature transformation on data sets characterized by small sample size and evident lag. Secondly, the SVR algorithm was adopted due to its robust ability to generalize and classify non-linearly in regression analysis of the data set. Lastly, GA was utilized to optimize the parameters of SVR, thereby enhancing the prediction accuracy of the model. Three data sets comprised of meteorological factors, air quality and the number of patients with respiratory diseases was utilized to conduct simulation and comparative experiments using the ISGS model. The experimental results demonstrate that the proposed ISGS model achieves a prediction accuracy of 93.65%, surpassing that of all other reference models. Furthermore, the model exhibits superior generalization capabilities and can effectively handle data sets with higher dimension and smaller sample size.

Key words: ISOMAP; SMOTE; GA; SVR; integrated model
Foundation Item(s): National Education Science Plan-Key project of the Ministry of Education (No.DLA190426)

1 引言

近年来,随着机器学习和数据挖掘在各个领域的广泛应用,单一算法模型已难以满足研究数据的多样化特性. 为解决这一问题,集成模型和组合模型由于其解决问题的高效性,深受学者们的广泛关注. 集成学习最早由Dasarathy和Sheela两位学者于1979年首次提出的[1],是一种将多个基础弱分类器通过不同的方式训练、集成新的学习方法,从而获取性能表现优于单个分类器的模型[2]. 而组合学习不仅限于分类器,其目标是结合多种机器学习模型来解决同一问题,通过信息互

收稿日期:2022-03-09;修回日期:2022-07-30;责任编辑:梅志强

补实现每个模型的优点最大化. 组合学习常用的研究思路包括串联结构和并行结构两种[3]. 文献[4]提出了一种基于百度指数的CEEMD-GRNN(Complementary Ensemble Empirical Mode Decomposition-Generalized Regression Neural Network)组合模型来研究HIV(Human Immunodeficiency Virus)感染病例数的周期以及非线性关系. 文献[5]在青少年健康评价研究中提出了基于决策树与BP(Back Propagation)神经网络的组合模型,得到了很好的评价效果. 文献[6]提出了一种基于集成树-梯度提升决策树的$PM_{2.5}$(Particulate Matter, 2.5)预测模型,提高了模型的泛化能力.

在回归预测模型的研究过程中发现,高维度、小样本数据集在训练中极度不稳定,训练过程中容易出现过拟合或欠拟合现象,难以实现良好的预测精度和泛化能力. 为解决这一问题,提出一种基于多种机器学习算法的组合模型ISGS. 该模型基于ISOMAP(Isometric Feature Mapping)算法、SMOTE(Synthetic Minority Oversampling Technique)算法、GA(Genetic Algorithm)算法和SVR(Support Vector Regression)算法,在对实验数据进行特征变换和新样本生成的基础上,对SVR算法进行参数优化,以提高基于单一SVR预测模型的泛化能力和预测精度.

2　相关研究

2.1　等距特征映射算法

等距特征映射作为一种非线性的降维算法,衍生于多维缩放(MultiDimensional Scaling, MDS)算法[7],能够对数据进行全局优化. 该算法在MDS算法的基础上引入邻域图,在保持降维后样本间距离不变的基础上,利用欧式距离和最短路径来度量高维空间中距离较近点、距离较远点之间的距离,从而实现对数据的降维保距[8].

2.2　少数类过采样技术

少数类过采样技术SMOTE算法是由Chawla等人于2002年提出的一种合成少数类数据的过抽样算法[9]. SMOTE算法的基本思路是分析训练数据的少数类样本后,利用人工合成出新样本,从而提高数据的样本容量,避免了一般过抽样算法的简单复制导致的模型过拟合问题. 由此被广泛应用于互联网、金融及医学等[10]领域的数据分析和数据挖掘[11]中.

2.3　遗传算法

遗传算法GA也称为进化算法,借鉴了达尔文的种群进化理论和基因遗传进化理论,通过模拟种群的自然淘汰和个体基因的遗传变异过程来达到搜索问题最优解的目的. 种群中的任一个体都是一个可行解,通过模仿生物的进化过程不断进行选择、交叉、变异,从而在解空间内自适应地搜索最优解[12].

2.4　支持向量回归

支持向量回归SVR是机器学习中用于研究回归问题的算法,具备支持向量机的特性,适合处理样本数有限或较少的数据集. 如图1所示,SVR的训练流程主要包括输入层、中间层和输出层三个部分[13]. SVR的核心思想是利用核函数实现特征空间的转换,通过线性函数来达到非线性回归的效果[14].

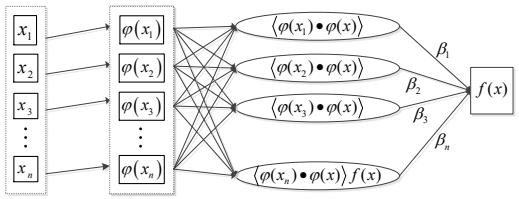


图1　SVR训练流程

3　ISGS模型算法

3.1　ISGS模型框架

出于对SVR算法的强非线性表达能力以及适合处理样本数有限或较少数据集两方面的考量,本文提出了基于4种机器学习算法的ISGS模型(如图2所示).

ISGS模型包含两部分内容. 第一部分是基于ISOMAP和SMOTE算法构建特征提取模型. 考虑到滞后效应对研究问题的影响,在对原始数据的缺失值进行填充后,利用特征提取实现实验数据的特征变换以及新样本生成. 经过特征提取后的实验数据进入第二部分. 模型的第二部分是基于GA-SVR的预测模型. 利用GA算法对训练样本进行编码,转换为具有特定数量并具有基因序列的种群. 通过不断地选择、交叉和变异,增强种族的适应性,进而得到SVR参数的最优解,以此来提高模型的泛化能力和预测精度.

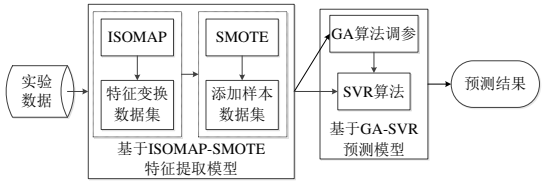


图2　ISGS模型框架图

3.2　模型算法

3.2.1　特征提取

ISOMAP算法采用最短路径距离构建变换矩阵,能够更好地保持数据空间中样本点的几何结构信息. 而SMOTE算法则是利用少数类样本在特征空间中的相似度,人工生成新样本,解决小样本、非平衡数据集的少数类数据的不均衡问题.

3.2.2 参数优化

GA算法通过模拟自然选择和自然遗传,将训练样本转换为每一个体代表每一个解的种群. 根据种群个体的适应度,按照如图3所示的流程进行操作,对个体不断进行优胜劣汰,直到求得最优解为止[15]. 其中,适应度是评价种群个体优劣的指标,常记为f,计算方法如式(1)所示的平均误差法,平均误差与个体适应度的变化趋势成反比.

$$f=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-y_i^*}{y_i}\right| \tag{1}$$

式中,y和y^*分别为预测值与实际值.

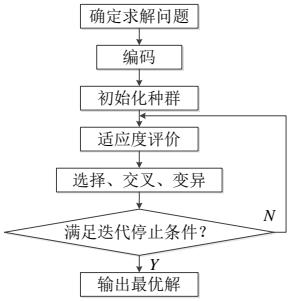


图3 GA算法流程图

在SVR常用的核函数中,相比于线性核函数和多项式核函数等[16],高斯径向基核函数(Radial Basis Function, RBF)能够处理自变量与因变量之间复杂的非线性问题,且计算相对更简单,参数更少、收敛域更宽[17]. 因此,本文在利用GA-SVR算法构建模型时,选择最为常用的RBF核函数,其数学关系表达式为:

$$K\left(x_1,x_2\right)=\exp\left(-\frac{\left\|x_1-x_2\right\|^2}{2g^2}\right) \tag{2}$$

式中,g为核参数.

基于RBF核函数的SVR模型,需要进行编码寻优的参数为决定函数离群点和拟合程度的最优惩罚函数c,以及影响样本数据子空间分布的复杂程度核参数g. 值得注意的是,当c过大时,函数的拟合程度高,泛化能力差.

3.2.3 算法流程

综上所述,本文提出的ISGS模型的算法流程如算法1所示.

算法1 ISGS模型算法

输入:N维原始数据,记为$X=\left\{x_i\middle|i=1,2,\cdots,n\right\}$,$X\in\mathbf{R}^N$,$n$为样本数量.

输出:预测结果.

Step1 基于ISOMAP-SMOTE算法特征提取.

(1) 构建邻域图$\boldsymbol{G}$:设定低维空间的维数d和近邻值k,以近邻点x_j与x_i间的欧式距离作为测量指标,记为$\boldsymbol{d}_{\boldsymbol{G}}(x_i,x_j)$,构建邻域图$\boldsymbol{G}$.

(2) 计算测地距离矩阵$\boldsymbol{D}_{\boldsymbol{M}}$:计算邻域图$\boldsymbol{G}$中任意两点间的最短路径,记为$\boldsymbol{d}_{\boldsymbol{G}}(x_i,x_j)$;近似流形$\boldsymbol{M}$上的测地距离,记为$\boldsymbol{d}_{\boldsymbol{M}}(x_i,x_j)$,得到距离矩阵$\boldsymbol{D}_{\boldsymbol{M}}$.

(3) 定义低维嵌入$\boldsymbol{H}$[18]:

$$\boldsymbol{H}=-\frac{1}{2}\left[\boldsymbol{I}-\frac{1}{n}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\right]\boldsymbol{D}_{\boldsymbol{M}}\left[\boldsymbol{I}-\frac{1}{n}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{3}$$

式中,$\boldsymbol{I}$为n阶单位阵;$\boldsymbol{e}$为元素均为1的n维列向量.

(4) 计算d维输出:对$\boldsymbol{H}$进行谱分解,取前$d\,(d\ll D)$个最大特征值,构成对角矩阵$\boldsymbol{V}$并计算其特征向量$\boldsymbol{U}$.

$$\boldsymbol{V}=\mathrm{Diag}\left(\lambda_1,\lambda_2,\cdots,\lambda_d\right) \tag{4}$$

$$\boldsymbol{U}=\left(\mu_1,\mu_2,\cdots,\mu_d\right) \tag{5}$$

(5) 计算降维输出结果$\boldsymbol{Y}$:

$$\boldsymbol{Y}=\sqrt{\boldsymbol{V}}\,\boldsymbol{U}^{\mathrm{T}} \tag{6}$$

(6) 计算少数类中任一样本y到其样本集中所有样本的欧氏距离,得到其k近邻.

(7) 根据样本集的不平衡比例设定采样倍率N,在少数类样本y的k近邻中随机选择若干个样本y_n.

(8) 根据y_n计算构建的新样本y_{new}:

$$y_{\mathrm{new}}=y+\mathrm{rand}\left(0,1\right)\times\left|y-y_n\right| \tag{7}$$

Step2 参数优化

(1) 初始化二进制编码,初始种群和最大迭代数,并初始化SVR算法的参数c,g.

(2) 将通过ISOMAP-SMOTE特征提取处理后的实验数据y_{new}读入SVR输入层,SVR根据输入的数据进行仿真训练.

(3) 将实验数据分为P个大小相似的互斥子集,利用5折交叉验证,计算测试集的均方误差BestMSE.

(4) 利用GA的基因算子对训练集样本进行选择、交叉和变异,编码产生新的c、g参数后,再次根据SVR输入层读入的数据进行仿真训练,从而使SVR具备在线学习的能力.

(5) 再次重复Step 2(4),得到新的BestMSE. 如果低于之前的BestMSE,那么将当前准确率赋值给BestMSE,否则,保持原值不变. 达到最大迭代次数或者最好预测精度,则输出Bestc、Bestg和BestMSE;否则,返回Step 2(5)继续运算.

(6) 利用Step 2(5)得出的最优参数,得到基于GA-SVR的预测结果,算法结束.

4 仿真实验

4.1 实验数据集

本文的模型实验数据来自国家人口健康科学数据中心数据仓储PHDA(https://www.ncmi.cn)的三个数据集:

(1)气象数据来源于2010年全国700个气象站的地面气象数据,从中提取出了北京市的气象因素资料数据,其中包括:气压(hPa)(平均本站气压PA,日最高本站气压PH,日最低本站气压PL);气温(℃)(日平均气温TA,日最高气温TH,日最低气温TL);湿度(%)

(平均相对湿度 HA,最小相对湿度 Hmin);日降水量 R(mm);风速(m/s)(平均风速 WA,最大风速(10 分钟平均风速)WH,极大风速 W_{max});日照时数 Sun(h)等 13 项因素.

(2)空气质量数据来源于环保部发布的 2010-01-01—2010-12-31 全国省会城市空气污染指数资料,从中提取出了北京市的空气污染指数 API(Air Pollution Index)($\mu g \cdot m^{-3}$). 两组数据中的少量缺失值使用均值法进行填充.

(3)呼吸系统发病人数统计数据来自北京三家医院 2010-01-01—2010-12-31 呼吸系统疾病日发病人数资料,数据包括按性别划分的呼吸系统疾病日发病人数.

4.2　基于 ISOMAP-SMOTE 算法特征提取

4.2.1　基于 ISOMAP 算法特征变换

由于气象因素和空气质量对发病人数的影响存在明显的滞后和累积效应[19],特引入各项变量的前 1 ~ 7 天的观测值作为滞后因子,因此整理后的实验数据包括 15 项变量的当天、前 1 ~ 7 天的观测值在内的共 120 个维度,来研究气象因素和空气质量与呼吸系统日发病人数之间的关系.

本文利用 ISOMAP 算法对加入滞后因子的实验数据进行特征变换后,数据维度由 120 转换成 15,样本量为 357.

4.2.2　基于 SMOTE 算法新样本生成

由于原始数据集的样本量小,模型难以充分训练、实现良好的泛化能力. 此外,考虑到呼吸系统发病人数受气象因素、空气质量等因素影响外,会随着时间序列的变化而有所差异,如图 4 所示. 2010 年北京市呼吸系统发病人数波动较大,在 2 月显著增加后逐渐回落,3 月降到最低,此后到 9 月底波动增加,12 月达到最大值. 由于数据量不充分,难以获得其他年份数据,为保证数据集的波动特性、使其数据量充足且相对平衡,利用 SMOTE 算法生成新样本集:设定采样倍率 N=5(模拟 2010 年前后近 5 年数据),新数据集样本量增加到 1 785 个,解决了原始数据集由于样本数量过少造成预测结果的过拟合而导致模型训练不充分等问题.

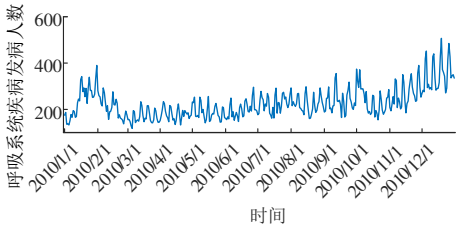


图 4　北京市 2010 年呼吸系统发病人数分布图

4.3　参数优化与预测结果分析

本文利用 GA 算法对基于 RBF 核函数的 SVR 模型的参数(c,g)进行在线优化. 在参数优化过程中,适应度函数的最小均方误差(Mean Square Error, MSE)设置为 5 折交叉验证,最大进化代数设置为 400,种群规模设置为默认值 20. 初始化完成后,GA 算法对预测模型中的 SVR 进行参数优化,参数寻优过程中的适应度变化曲线如图 5 所示,反映了每一代群体的最佳适应度以及平均适应度的进化过程.

由图 5 可知,SVR 的最佳参数值为:c=2.928,g=6.231 3,此时训练集的 MSE=0.005 631 5. 根据适应度曲线可以看出,模型在训练后期,其最佳适应度基本在[0.005,0.013]区间内保持小幅度波动,表明了训练样本的多样性,同时还具有较好的收敛性.

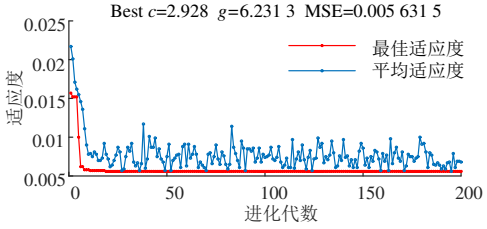


图 5　GA 算法优化参数适应度曲线变化图

如图 6 所示,将最优的 Bestc、Bestg 带入 SVR 模型,得到的训练集预测结果,均方误差 MSE=0.002 150 5,拟合度 R^2 较好,接近于 1. 随后,将测试集数据输入到 GA-SVR 预测模型中,如图 7 所示,测试样本得到的均方误差 MSE=0.002 150 5,预测结果的拟合度 R^2=93.65%,预测结果和真实值基本一致.

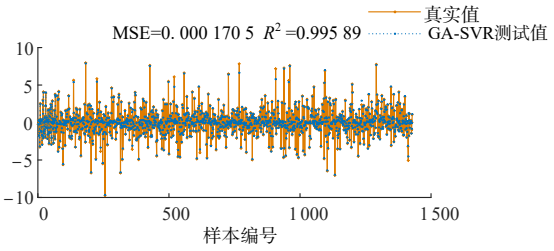


图 6　GA-SVR 训练集预测结果对比

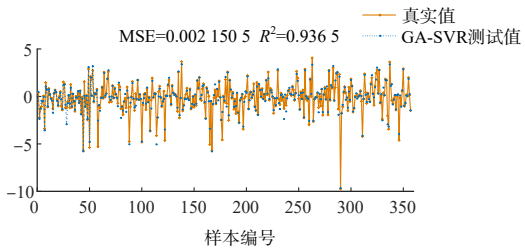


图 7　GA-SVR 测试集预测结果对比

5 对比实验

为了更好地验证所提出的ISGS模型的有效性，设置了2组对比实验：

(1)采用特征提取模型与未使用该模型的结果对比.

(2)采用GA优化的SVR与使用传统交叉验证(CV)获取参数的SVR的对比.

5.1 特征提取模型使用前后对比分析

特征提取模型使用前后对比结果如表1所示，可以看出使用ISOMAP-SMOTE特征提取模型后，基于SVR和GA-SVR的2种预测模型的预测精度均有提升，更加接近真实值.

表1 特征提取前后GA-SVR模型预测结果对比

模型		预测结果	
		原始数据	ISOMAP-SMOTE特征提取数据
GA-SVR	MSE	0.085 512	0.002 150 5
	R^2	32.141%	93.65%
SVR	MSE	0.126 57	0.005 989 8
	R^2	16.969%	83.481%

5.2 模型参数优化对比分析

为验证通过GA算法优化参数的有效性，本文使用单一SVR算法，利用传统交叉验证方法寻找最优的(c, g)，经过仿真训练后，测试集的MSE=0.005 989 8，R^2=83.481%，如图8所示. 单一SVR模型的预测结果与使用GA算法优化参数后的GA-SVR模型的预测结果对比如表2所示.

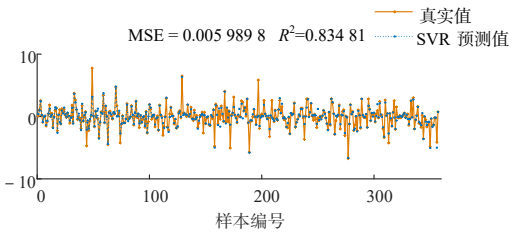


图8 SVR测试集预测结果对比

表2 模型参数优化对比

预测结果	SVR	GA-SVR
MSE	0.005 989 8	0.002 150 5
R^2	83.481%	93.65%

由表2可知，运用GA算法进行参数优化后，GA-SVR预测模型的MSE值有所降低，预测精度R^2为93.65%，与单一SVR预测模型的83.481%相比，预测精度提高了约10%，表明使用GA算法优化的GA-SVR预测模型的预测效果更佳.

6 结论

本文构建了一种处理滞后效应和累积效应明显、小样本数据集的ISGS模型. 利用ISOMAP和SMOTE算法构建的特征提取模型和基于GA-SVR的预测模型实现ISGS模型的构建，在对实验数据进行特征提取的同时，提高了模型的泛化能力和预测精度. 综合研究结果表明，经过ISOMAP-SMOTE算法进行特征提取后，能够极大程度提高模型的预测精度；与单一的SVR模型相比，GA-SVR模型具有较高的预测精度，预测精度由83.481%提升到93.65%，证明本文建立的组合模型具备一定的可行性和有效性，为未来在特征工程和遗传算法算子优化，提高模型的鲁棒性和泛化能力等方面提供了研究基础.

参考文献

[1] DIETTERICH T G. Ensemble methods in machine learning[C]//Proceedings of the First International Workshop on Multiple Classifier Systems. Berlin: Springer-Verlag, 2000: 1-15.

[2] 崔鸿雁, 徐帅, 张利锋, 等. 机器学习中的特征选择方法研究及展望[J]. 北京邮电大学学报, 2018, 41(1): 1-12.
CUI H Y, X U, ZHANG L F, et al. The key techniques and future vision of feature selection in machine learning[J]. Journal of Beijing University of Posts and Telecommunications, 2018, 41(1): 1-12. (in Chinese)

[3] 牛晓健, 凌飞. 基于组合学习的个人信用风险评估模型研究[J]. 复旦学报(自然科学版), 2021, 60(6): 703-719.
NIU X J, LING F. Study on personal credit risk assessment model based on hybrid learning[J]. Journal of Fudan University (Natural Science), 2021, 60(6): 703-719. (in Chinese)

[4] 魏麟, 朱素玲, 胡晓斌. 基于CEEMD-GRNN组合模型的HIV感染病例数预测[J]. 现代预防医学, 2022, 49(6): 969-974.
WEI L, ZHU S L, HU X B. Prediction of HIV infection cases based on CEEMD-GRNN model[J]. Modern Preventive Medicine, 2022, 49(6): 969-974. (in Chinese)

[5] PENG Y, XU J, DING X X, et al. Health assessment of young students based on decision Tree-BP model[J]. Journal of Nonlinear and Convex Analysis, 2019, 20(5): 977-986.

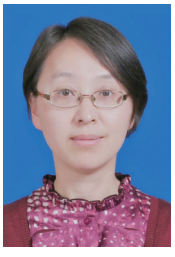
[6] 彭岩, 赵梓如, 吴婷娴, 等. PM2.5浓度预测与影响因素分析[J]. 北京邮电大学学报, 2019, 42(6): 162-169.
PENG Y, ZHAO Z R, WU T X, et al. Prediction of PM2.5 concentration based on ensemble learning[J]. Journal of Beijing University of Posts and Telecommunications,

2019, 42(6): 162-169. (in Chinese)
[7] TENENBAUM J B, DE SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.
[8] 石陆魁, 郭林林, 房子哲, 等. 基于Spark的并行ISOMAP算法[J]. 中国科学技术大学学报, 2019, 49(10): 842-850.
SHI L K, GUO L L, FANG Z Z, et al. Parallel ISOMAP algorithm based on Spark[J]. Journal of University of Science and Technology of China, 2019, 49(10): 842-850. (in Chinese)
[9] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
[10] HE H B, GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.
[11] 钟龙申, 高学军, 王振友. 一种新的基于K-means改进SMOTE算法在不平衡数据集分类中的应用[J]. 数学的实践与认识, 2015, 45(19): 198-206.
ZHONG L S, GAO X J, WANG Z Y. A new kind of improving SOMTE algorithm based on K-means in imbalanced datasets[J]. Mathematics in Practice and Theory, 2015, 45(19): 198-206. (in Chinese)
[12] 史耀凡, 栾元重, 于水, 等. 基于PCA-GA-SVM模型的地表下沉系数预测[J]. 矿业研究与开发, 2022, 42(2): 65-69.
SHI Y F, LUAN Y Z, YU S, et al. Prediction of surface subsidence coefficient based on PCA-GA-SVM model[J]. Mining Research and Development, 2022, 42(2): 65-69. (in Chinese)
[13] 张成成, 陈求稳, 徐强, 等. 基于支持向量机的太湖梅梁湾叶绿素*a*浓度预测模型[J]. 环境科学学报, 2013, 33(10): 2856-2861.
ZHANG C C, CHEN Q W, XU Q, et al. A chlorophyll-a prediction model for meiliang bay of taihu based on support vector machine[J]. Acta Scientiae Circumstantiae, 2013, 33(10): 2856-2861. (in Chinese)
[14] CHEN Y W, LIN C J. Combining SVMS with various feature selection strategies[M]//Feature Extraction. Berlin: Springer , 2008: 315-324.
[15] WU X G, ZHU Y P. A mixed-encoding genetic algorithm with beam constraint for conformal radiotherapy treatment planning[J]. Medical Physics, 2000, 27(11): 2508-2516.
[16] 石怀涛, 赵纪宗, 宋文丽, 等. 基于人工蜂群优化核主元分析故障检测方法[J]. 控制工程, 2018, 25(9): 1686-1691.
SHI H T, ZHAO J Z, SONG W L, et al. Fault detection method with kernel principal component analysis based on artificial bee colony optimization[J]. Control Engineering of China, 2018, 25(9): 1686-1691. (in Chinese)
[17] KEERTHI S S, LIN C J. Asymptotic behaviors of support vector machines with Gaussian kernel[J]. Neural Computation, 2003, 15(7): 1667-1689.
[18] 屈太国, 蔡自兴. 基于分而治之的多维标度算法[J]. 模式识别与人工智能, 2014, 27(11): 961-969.
QU T G, CAI Z X. A divide-and-conquer based multidimensional scaling algorithm[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(11): 961-969. (in Chinese)
[19] 李娟, 张志薇, 于庚康, 等. 气象要素对南京市呼吸系统疾病的影响研究[J]. 气象科学, 2017, 37(3): 409-415.
LI J, ZHANG Z W, YU G K, et al. Impact of meteorological factors on respiratory diseases in Nanjing[J]. Journal of the Meteorological Sciences, 2017, 37(3): 409-415. (in Chinese)

作者简介

冯婷婷　女,1992年出生,天津人,硕士研究生. 主要研究方向为公共管理信息化理论与技术.
E-mail: fengtt0702@163.com

彭　岩　女,1967年出生,重庆人,博士,教授. 主要研究方向为大数据分析与数据挖掘.
E-mail: pengyan@cnu.edu.cn



王　洁(通讯作者)　女,1977年出生,湖北黄石人,博士,副教授. 主要研究方向为数据挖掘、机器学习.
E-mail:wangjie@cnu.edu.cn