湖南农业大学学报(自然科学版)　2020, 46(4):495–500. DOI:10.13331/j.cnki.jhau.2020.04.018
Journal of Hunan Agricultural University (Natural Sciences)

引用格式：
王文，饶元，李绍稳，Arthur GENIS．基于预测模型的异常农情数据在线检测方法的研究[J]．湖南农业大学学报(自然科学版)，2020，46(4)：495–500．
WANG W, RAO Y, LI S W, Arthur GENIS. Research on online detection method for the abnormal agricultural data based on prediction model[J]. Journal of Hunan Agricultural University(Natural Sciences), 2020, 46(4): 495–500.
投稿网址：http://xb.hunau.edu.cn

基于预测模型的异常农情数据在线检测方法的研究

王文[1]，饶元[1*]，李绍稳[1]，Arthur GENIS[2]

(1.安徽农业大学信息与计算机学院，安徽 合肥 230036；2.Katif 沿海沙漠开发研究中心，以色列 内提沃特 8771002)

摘　要：为保证农业物联网传感器的数据感知质量，构建了基于滑动窗口和预测模型(支持向量回归、K 近邻、梯度提升回归和随机森林)的异常农情数据在线检测框架，提出了基于数据特征的滑动窗口尺寸计算方法，运用熵权逼近最优排序法评价预测模型适用性。采用羊圈环境数据(空气温度、相对湿度、CO_2 和 H_2S 体积分数)进行试验，结果表明，滑动窗口尺寸计算方法优于仅基于采样间隔和特征周期的计算方法；模型预测误差与其异常检测性能负相关，且对误检率影响更大；支持向量回归模型对空气温度和相对湿度异常数据检测适用性最好，贴近度达 0.8 以上，梯度提升回归和 K 近邻模型分别对 CO_2 和 H_2S 体积分数异常数据检测适用性较优，两者贴近度均在 0.6 左右。

关　键　词：农业物联网；异常数据；在线检测；预测模型

中图分类号：TP391　　文献标志码：A　　文章编号：1007–1032(2020)04–0495–06

Research on online detection method for the abnormal agricultural data based on prediction model

WANG Wen[1], RAO Yuan[1*], LI Shaowen[1], Arthur GENIS[2]

(1.College of Information and Computer Sciences, Anhui Agricultural University, Hefei China, 230036; 2.Katif Research Center for Coastal Deserts Development, Netivot Israel, 8771002)

Abstract: In order to guarantee the quality of perceptional data, an online detection framework for the abnormal agricultural data is constructed based on the sliding window and the prediction models, which including support vector regression, K-nearest neighbor, gradient boosting regression and random forest. The calculation method of the sliding window size is proposed based on data features. The applicability of the prediction models is evaluated by using entropy weight TOPSIS. Through the sheepfold's monitoring data of the air temperature, the relative humidity, and the CO_2 and H_2S volume fractions, it is demonstrated that the proposed calculation method of sliding window size is superior to the calculation method simply based on the sampling interval and characteristic period. The prediction errors of these models are negatively correlated with the abnormal detection performance and could impose significant influence on false positive rate. Support vector regression model is the most appropriate candidate for detecting the abnormal data in air temperature and relative humidity with the close degree greater than 0.8, whereas the most appropriate candidates for dealing with CO_2 and H_2S volume fractions are gradient boosting regression model and K nearest neighbor model, both of them with the close degrees of 0.6.

Keywords: agricultural internet of things; abnormal data; online detection; prediction model

收稿日期：2019–11–21　　修回日期：2020–05–18

基金项目：农业部引进国际先进农业科学技术“948”项目(2015–Z44、2016–X34)；安徽省自然科学基金项目(1608085QF126)；安徽省重点研究和开发计划面上攻关项目(1804a07020108，201904a06020056)；安徽农业大学省级大创项目(201910364263)

作者简介：王文(1996—)，男，安徽淮北人，硕士研究生，主要从事农业物联网技术研究，wangwen0815@foxmail.com；*通信作者，饶元，博士，教授，主要从事农业信息技术研究，raoyuan@ahau.edu.cn

农业物联网中，传感器按照时间序列连续采集作物生长、畜禽生长和环境指标等信息，以数据流的形式传输至数据中心，具有显著周期性、实时性、无穷性等特征[1–2]。农业物联网设备受制造技术、工艺与成本以及网络传输的影响，在数据收集过程中不可避免地会产生异常数据[3–4]，使得数据质量急剧下降，无法保证物联网设备的智能调控和数据的有效分析；因此，实时高效地检测出农业物联网中的异常数据，对于农业物联网的管理与决策具有重要意义[5–6]。

异常数据在线检测通常采用预测技术对实时数据流进行分析，通过设置滑动窗口大小确定预测模型的训练数据[7–8]。苑进等[9]采用自回归高斯过程模型计算的预测误差带进行大棚内空气温度和相对湿度的异常值检测，但高斯过程回归模型受初值和协方差函数影响较大；段青玲等[10]采用基于滑动窗口的支持向量回归(SVR)预测畜禽养殖物联网数据流，通过判断实际测量值是否在给定的置信区间中来检测异常值，能达到一定的检测率，但滑动窗口尺寸的判定忽略了数据本身的特征；XIE 等[11]提出了一种基于超网格直觉的 K 近邻(KNN)的无线传感器网络异常数据检测方案，能自适应估计模型参数，可应用于多种场景，但未涉及异常数据的修正；梯度提升回归(GBR)和随机森林(RF)也被广泛应用于空气质量等数据的预测[12–13]，但运行时间较长，且其异常检测性能研究尚未涉及。

滑动窗口的大小是影响在线检测精度与运行时间的重要因素。农业物联网中数据特征与采样间隔往往存在较大的差异，如何有效地确定滑动窗口尺寸一直面临着挑战。此外，由于各模型的异常检测能力、运算复杂度各异，其适用性也不明确。笔者构建了异常农情数据在线检测框架，根据农情数据特征与采样间隔探索了滑动窗口尺寸计算方法，采用支持向量回归、K 近邻、梯度提升回归和随机森林 4 种数据预测模型，对羊圈中 4 类农情数据(温度、相对湿度、CO_2 和 H_2S 体积分数)进行异常检测，根据预测模型的检测率、误检率和运行时间等评估参数，运用熵权最优排序法对 4 种预测模型开展适用性评价，以期为实现农业物联网高质量数据收集提供参考。

1 数据来源

数据来自于安徽省长丰县双墩镇合肥安谷农业有限公司羊圈监测点。部署的传感器节点不间断地采集空气温度和相对湿度、CO_2 和 H_2S 体积分数等数据，采样间隔为 5 min。截取 2019 年 3 月 11 日至 22 日共 12 d(288 h，3 456 个采样点)的观测数据，其中，空气温度和相对湿度分别有 21 和 24 个异常值，CO_2 体积分数有 29 个异常值，H_2S 体积分数有 45 个异常值。为便于开展检测性能对比分析，在 4 类数据中增加了部分人工噪声点，使得各数据异常点数为 60 个[14]。增加人工噪声后的数据如图 1 所示。4 类数据的范围、变化特征存在较大差异：

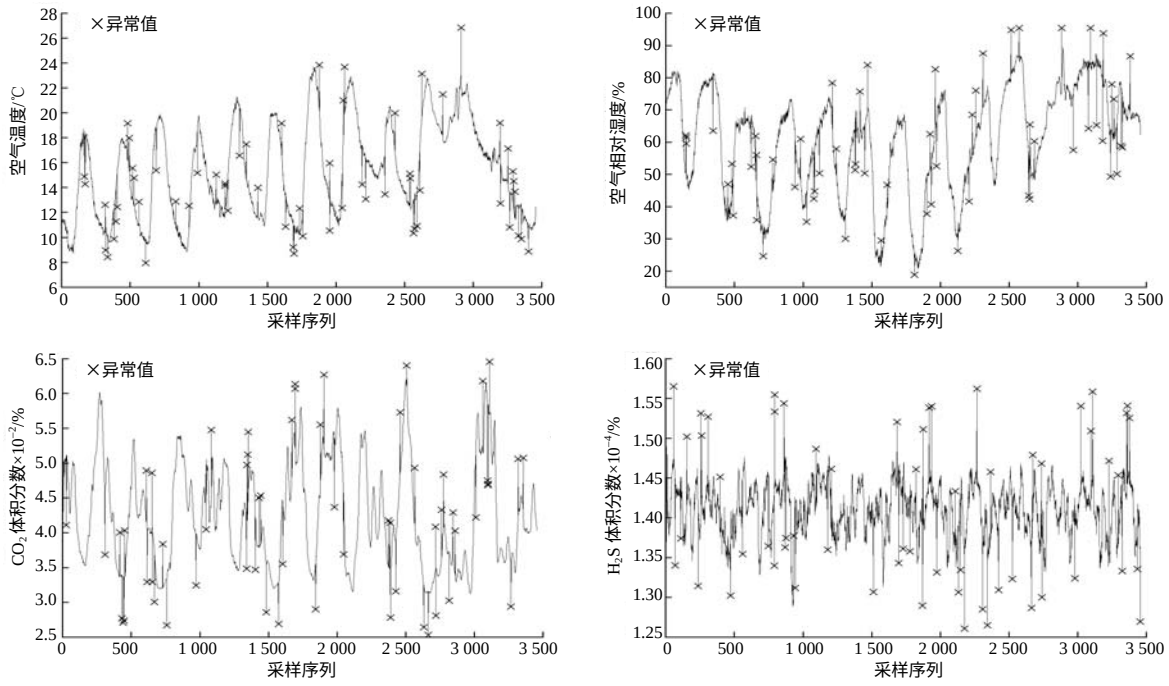

图 1 添加异常值后羊圈环境的监测数据

Fig.1 Sheepfold's monitoring data with the added abnormal data

空气温度和相对湿度具有明显的周期性分布规律，且相邻数据间差异较小；CO_2 体积分数具有强周期性特征，相邻数据间波动较大；H_2S 体积分数的周期性特征较弱，且相邻数据间波动较大。

2　异常农情数据在线检测框架的建立

异常农情数据在线检测框架如图 2 所示。基于数据采样间隔、离散程度和周期特征，提出滑动窗口尺寸计算方法：首先将特征周期 T 内所采集的数据进行归一化处理，再根据归一化后数据的离散程度、采样间隔等计算滑动窗口尺寸 l。

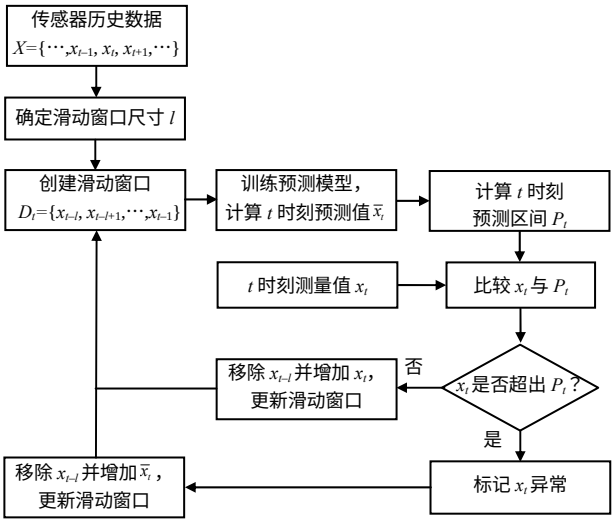

图 2　异常农情数据在线检测框架

Fig.2　Online detection framework for the abnormal agricultural data

$$l=\left\lceil \frac{S\times T}{\Delta t}\mathrm{e}^{-\beta\ln\sqrt{T/\Delta t}} \right\rceil \tag{1}$$

式中：Δt 为传感器数据的采样间隔；β 为支持度衰减因子；S 为特征周期 T 内数据归一化后的标准差。

基于预测模型的异常农情数据在线检测流程包括以下步骤。

1) 基于数据特征和采样间隔确定滑动窗口尺寸 l；

2) 根据 l，选择 t 时刻所在采样序列之前 l 个测量值创建滑动窗口数据集 $D_t\{x_{t-l},x_{t-l+1},...,x_{t-1}\}$，$x_{t-l}$ 为 t 时刻之前的第 l 个测量值；

3) 训练预测模型，计算 t 时刻的预测值 $\bar{x}_t$；

4) 根据滑动窗口数据集、$\bar{x}_t$，采用学生 t 分布概率分布函数，求出 t 时刻的预测区间 P_t [15]。假设测量值落入预测区间的置信水平为 p=100×(1–α)，则预测区间可表示为：

$$P_t=\bar{x}_t\pm t_{\alpha/2,l\text{-}1}S\sqrt{1+1/l} \tag{2}$$

其中：P_t 为 t 时刻的预测区间；$t_{\alpha/2,\,l\text{-}1}$ 为 p 百分位数自由度为(l–1)的符合学生 t 分布概率分布函数。

5) 根据实际测量值 x_t 是否超出预测区间 P_t 来判断是否异常，如异常，则剔除异常数据，并采用预测值 $\bar{x}_t$ 填补。

6) 更新滑动窗口数据集，重复第 2 至第 5 步骤。

采用支持向量回归(SVR)[10]、K 最近邻(KNN)[11]、梯度提升回归(GBR)[12]和随机森林(RF)[13]4 种算法作为数据预测模型对农情数据流中的异常数据进行检测，选择 t 时刻之前的滑动窗口内的 l 个测量值作为输入值，利用预测模型对 t 时刻传感器测量值进行预测，输出值为 t 时刻的预测值。通过试验确定模型超参数：SVR 模型采用径向基核函数，惩罚系数 C 为 1.0；KNN 模型中 K 值为 5；GBR 模型基分类器数为 6；RF 模型基分类器数为 3，最大深度为 3。

采用熵权法(TOPSIS)[16]评价预测模型的适用性：确定评价指标权重后，构建加权规范化决策矩阵，再确定预测模型的最优、最劣指标向量，最后求出评价目标与最优指标组合方案的贴近度。

3　试验验证

采用 Python 3.6 语言编程、Windows 10 操作系统。硬件平台为配置 IntelCorei5–2400CPU、8GB 内存的台式计算机。为便于对比各模型预测误差，将数据归一化后再进行异常检测处理与分析，结果为 10 次试验的平均值。采用均方根误差度量各模型的预测误差。通过计算异常数据检测率、误检率对各预测模型在 95%置信水平上的异常检测性能进行评估[10]。

3.1　滑动窗口尺寸的确定

由图 1 可知，4 类监测数据变化的特征周期 T 均为 24 h，依据文献[10]，β 取 0.5，采样间隔 Δt 为 5 min。基于数据采样间隔、离散程度和特征周期，空气温度、空气相对湿度、CO_2 和 H_2S 体积分数数据的滑动窗口尺寸分别为 23、22、18 和 10，均小于采用 SW–SVR 方法[10]的滑动窗口尺寸为 37 时的计算结果。

以 SVR 模型为例，图 3 表示的是 4 种数据采

用 SW–SVR 方法和笔者建立的方法计算的窗口尺寸预测误差和运行时间。在模型预测误差方面，采用窗口尺寸时各模型的预测误差均小于 SW–SVR 方法。此外，较小的窗口尺寸降低了模型训练时长，使得各模型具有较短的运行时间。为验证式(1)得出窗口尺寸的合理性，评估滑动窗口尺寸分别为 10、15、20、25、30、35 和 40 时 SVR 模型的预测误差。图 4 表明，空气温度和相对湿度数据中，窗口尺寸位于[20,25]时预测误差最低，CO_2 数据在窗口尺寸位于[15,20]时预测误差最小，H_2S 数据在窗口尺寸为 10 时具有最小的预测误差，与式(1)得出的窗口尺寸基本吻合。

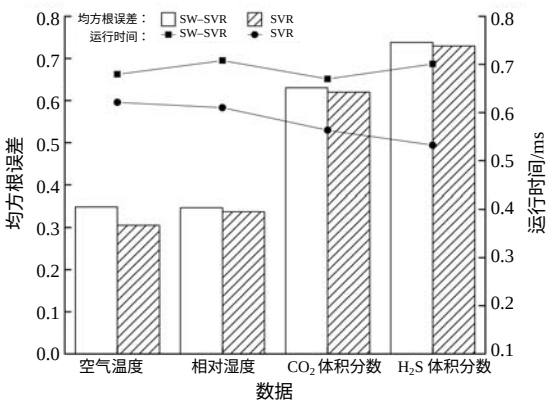

图 3 SVR 模型的对预测误差和运行时间

Fig.3 SVR model prediction error and computational time

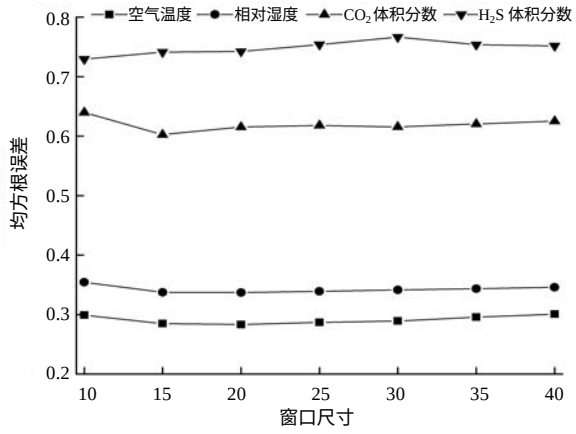

图 4 不同窗口尺寸下的模型预测误差

Fig.4 Prediction error of the models with different window sizes

结合 KNN、GBR 和 RF 模型，进一步考察提出滑动窗口尺寸计算方法的合理性。对于相同数据，同一预测模型的预测误差较为接近，且采用式(1)的滑动窗口尺寸整体上略优于 SW–SVR。KNN 模型的评估结果与 SVR 相似，空气温度和相对湿度数据中滑动窗口尺寸在 20 左右时预测误差最小，CO_2 数据在窗口尺寸为 15 至 20 之间预测误差最小，H_2S 数据误差随着窗口尺寸的增大而增大。GBR 和 RF 模型，空气温度和相对湿度数据预测误差随窗口尺寸同步增大。对于波动性较强的 CO_2 与 H_2S 数据最优窗口尺寸与式(1)较好吻合。总之，相比于 SW–SVR 的窗口尺寸，笔者提出的窗口尺寸计算方法能够有效降低预测误差和运行时间。

3.2 预测模型的异常检测性能

4 种预测模型对空气温度、空气相对湿度、CO_2 和 H_2S 体积分数数据中异常值的检测效果列于表 1。各模型的异常检测性能与数据自身的特征密切相关。对于空气温度和相对湿度数据，4 种模型的异常数据检测率较为接近，均达到了 95%以上，但不同模型间误检率差异较大，SVR 和 GBR 模型的误检率相近，且仅为 KNN 和 RF 模型误检率的 1/8 至 1/10。随着数据波动性的增强，4 种模型检测率均呈下降趋势。对于波动性较大的 CO_2 和 H_2S 数据，KNN 和 GBR 模型检测率和误检率均优于 SVR 和 RF 模型。

表 1 预测模型对异常值的检测率与误检率

Table 1 TPR and FPR of anomaly data with different models

模型	检测率/%				误检率/%			
	空气温度	空气相对湿度	CO_2 体积分数	H_2S 体积分数	空气温度	空气相对湿度	CO_2 体积分数	H_2S 体积分数
SVR	97.50	96.52	93.33	94.17	0.42	0.18	4.47	3.99
KNN	96.65	97.50	95.83	95.02	4.89	3.24	1.80	1.95
GBR	96.77	98.33	96.89	95.01	0.33	0.21	0.06	0.63
RF	95.83	96.37	94.17	93.33	2.39	2.85	2.70	3.15

各模型的预测误差与运行时间如图 5 所示。对于空气温度和相对湿度数据，SVR 和 GBR 模型的预测误差较为接近，且略优于 KNN 和 RF 模型。对于波动性较大的数据，4 种模型预测误差间的差异增大；整体而言，KNN 和 GBR 模型较为接近且优于 SVR 和 RF 模型。综上可得，模型的异常检测性

能与其预测误差呈负相关，且模型的预测能力对误检率影响更大，尤其对于波动性大的数据的误检率有着更大的影响。

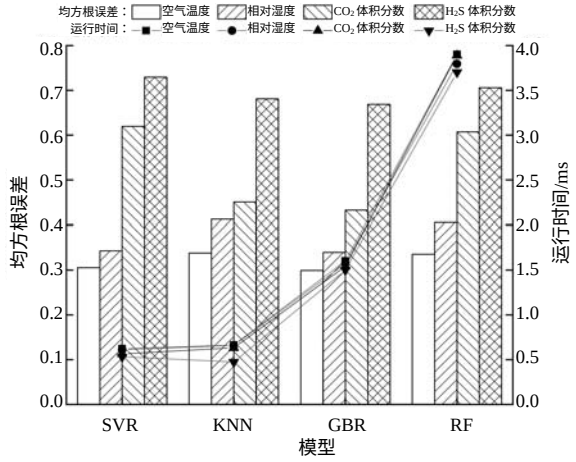

图 5　预测模型的预测误差与运行时间

Fig.5　Prediction error and computational time for the different models

在空气温度和相对湿度数据中，数据周期性较强且趋势较为平滑，4 种模型的预测误差较低且相差不大，模型的异常值检测灵敏度较高。相应地，模型间微小的预测误差差异会导致较大的误检率差异，如 SVR 和 GBR 模型的误检率远低于 KNN 和 RF 模型。在波动性较大的 CO_2 和 H_2S 数据中，4 种模型的预测误差均升高且差异增大，引起异常值检测敏感度下降，进而造成异常检测性能下降。此外，模型间的运行时间存在差异，SVR 和 KNN 模型的运行时间较为接近，分别为 GBR 和 RF 模型的约 1/3 和 1/6。

3.3　模型适用性评估

4 种预测模型的贴近度如图 6 所示。结果表明，检测空气温度和相对湿度数据的异常值时，模型优先级由高到低分别为 SVR、GBR、KNN 和 RF。检测 CO_2 数据的异常值时，模型优先级由高到低分别为 GBR、SVR、KNN 和 RF；检测 H_2S 数据的异常值时，模型优先级由高到低分别为 KNN、GBR、SVR 和 RF。这是因为优先级由预测模型的运行时间、异常值检测率和误检率共同决定的。对于空气温度和相对湿度数据，SVR 模型的异常值检测率与 KNN 和 RF 模型较为接近，但其误检率、运行时间均明显优于后两者；尽管 GBR 与 SVR 模型的检测性能较为接近，但 GBR 的运行时间是 SVR 的 3 倍，因此 SVR 模型贴近度高达 0.8 以上。对于波动性较强的 CO_2 和 H_2S 数据，SVR 模型检测性能的劣化导致其贴近度下降。对于 CO_2 数据检测，虽然 GBR 模型运行时间较长，但其优异的检测性能使得贴近度约为 0.6，仍然高于其他模型；对 H_2S 数据检测，由于 KNN 和 GBR 模型间的检测性能差距有所缩小，使得具有较短运行时间的 KNN 模型贴近度优于后者。

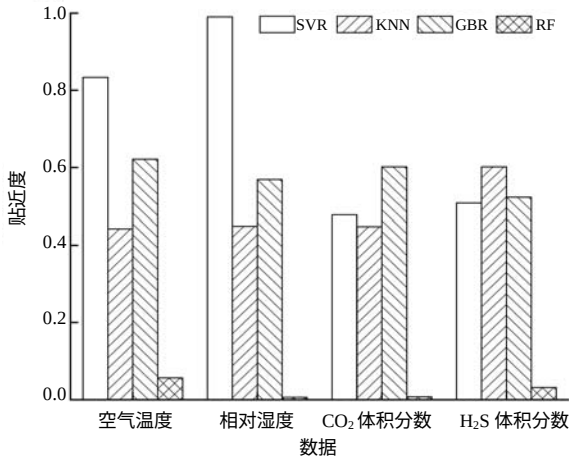

图 6　4 种预测模型的贴近度

Fig.6　Applicability evaluation results of four models

RF 模型缺乏优势，运行时间较长，并不适合于异常农情数据的在线检测。对于周期性强、波动性小的数据，SVR 模型能够在检测率、误检率和运行时间等方面取得良好的折衷，故是最优的方案。对于波动性大的农情数据，KNN 和 GBR 模型均可以考虑。由于实际应用中误报警往往会影响用户对自动检测系统的接受程度[17]，因而在对波动性较大的数据进行异常检测时，推荐从 KNN 和 GBR 模型中选择；若硬件运算资源较为充沛，优先选择 GBR 模型，否则 KNN 模型更能满足异常数据在线检测性能和运算效率的要求。

4　结论

本研究结果表明，基于数据特征计算滑动窗口尺寸，能够有效降低模型预测误差与运行时间。模型预测误差与其异常检测性能呈负相关，尤其是对于波动性较大的数据误检率影响更为显著。模型适用性与模型自身特征和数据对象有关。对于具有较低波动性数据的异常值检测，SVR 模型是最佳方案；而对于波动性较强的数据，KNN 和 GBR 模型

更为合适；若硬件运算资源充足，宜选择 GBR 模型，否则应考虑 KNN 模型。

参考文献：

[1] 王嘉宁，牛新涛，徐子明，等．基于无线传感器网络的温室 CO_2 浓度监控系统[J]．农业机械学报，2017，48(7)：280–295．
WANG J N，NIU X T，XU Z M，et al．Monitoring system for CO_2 concentration in greenhouse based on wireless sensor network[J]．Transactions of the Chinese Society for Agricultural Machinery，2017，48(7)：280–295．

[2] 白士宝，滕光辉，杜晓冬，等．基于 LabVIEW 平台的蛋鸡舍环境舒适度实时监测系统设计与实现[J]．农业工程学报，2017，33(15)：237–244．
BAI S B，TENG G H，DU X D，et al．Design and implementation on real-time monitoring system of laying hens environmental comfort based on LabVIEW[J]. Transactions of the Chinese Society of Agricultural Engineering，2017，33(15)：237–244．

[3] 李亮斌，姜晟，王卫星，等．基于无线传感器网络的农村供水厂水质监测节点的设计[J]．湖南农业大学学报(自然科学版)，2016，42(2)：212–216．
LI L B，JIANG S，WANG W X，et al．Design of wireless sensor network node for monitoring water quality of rural water supply plant[J]．Journal of Hunan Agricultural University(Natural Sciences)，2016，42(2)：212–216．

[4] ZHANG M，LI X，WANG L L．An adaptive outlier detection and processing approach towards time series sensor data[J]．IEEE Access，2019，7：175192–175212．

[5] NOSHAD Z，JAVAID N，SABA T，et al．Fault detection in wireless sensor networks through the random forest classifier[J]．Sensors，2019，19(7)：1568–1588．

[6] YU T Q，WANG X B，ABDALLAH S，Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems[J]．IEEE Internet of Things，2017，4(6)：2207–2216．

[7] REMI D，MAURIZIO F，PIETRO M，et al．A comparative evaluation of outlier detection algorithms：experiments and analyses[J]．Pattern Recognition，2018，74：406–421．

[8] TITOUNA C，FARID N，KHOKHAR A．DODS：a distributed outlier detection scheme for wireless sensor networks[J]．Computer Networks，2019，161：93–101．

[9] 苑进，胡敏，WANG K，等．基于高斯过程建模的物联网数据不确定性度量与预测[J]．农业机械学报，2015，46(5)：265–272．
YUAN J，HU M，WANG K，et al．Uncertainty measurement and prediction of IoT data based on gaussian process modeling[J]．Transactions of the Chinese Society for Agricultural Machinery，2015，46(5)：265–272．

[10] 段青玲，肖晓琰，刘怡然，等．基于 SW–SVR 的畜禽养殖物联网异常数据实时检测方法[J]．农业机械学报，2017，48(8)：159–165．
DUAN Q L，XIAO X Y，LIU Y R，et al．Anomaly data real-time detection method of livestock breeding internet of things based on SW-SVR[J]．Transactions of the Chinese Society for Agricultural Machinery，2017，48(8)：159–165．

[11] XIE M，HU J K，SONG H，et al．Scalable hypergrid k-NN-based online anomaly detection in wireless sensor networks[J]．IEEE Transactions on Parallel and Distributed Systems，2013，24(8)：1661–1670．

[12] 杨正理，史文，陈海霞，等．大数据背景下采用互信息与随机森林算法的空气质量预测[J]．环境工程，2019，37(3)：180–185．
YANG Z L，SHI W，CHEN H X，et al．Air quality forecasting with mutual information and random forests based on big data[J]．Environmental Engineering，2019，37(3)：180–185．

[13] KEPRATE A，RATNAYAKE R C．Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping[C]//Proceeding of IEEE International Conference on Industrial Engineering and Engineering Management．Piscataway：IEEE Press，2017：1331–1336．

[14] ZIDI S，MOULAHI T，ALAYA B．Fault detection in wireless sensor networks through SVM classifier[J]. IEEE Sensors Journal，2017，18(1)：340–347．

[15] RAO Y，ZHAO G，WANG W，et al．Adaptive data acquisition with energy efficiency and critical-sensing guarantee for wireless sensor networks[J]．Sensors，2019，19(12)：2654．

[16] CHEN P Y．Effects of normalization on the entropy-based TOPSIS method[J]．Expert Systems with Applications，2019，136：33– 41．

[17] TIM V D G，STEPHANIE V W，ANNELIES V N，et al. Supporting the development and adoption of automatic lameness detection systems in dairy cattle：effect of system cost and performance on potential market shares[J]. Animals，2017，7(10)：77–92．

责任编辑：罗慧敏
英文编辑：吴志立