

农业机械作业大数据清洗方法与试验优化

苑严伟 徐 玲 冀福华 郭大方 安 飒 牛 康

(中国农业机械化科学研究院土壤植物机器系统技术国家重点实验室, 北京 100083)

摘要: 针对农业机械大数据平台中,已有数据清洗算法不适用于大规模、多源异构、高维度和强时空相关实时数据的问题,分析了复杂田间环境下农机作业数据异常来源及特征,研究了异常数据检测及修正技术,提出一种基于滑动窗口机制的农机作业数据在线清洗方法。该方法基于方差约束原则识别异常数据,基于最小变动原则生成候选修正数据,基于数据时间相关性通过 AR、ARX 模型迭代优化得到最终修复值,依托 Flink 分布式计算平台,从而适应农机数据吞吐量大、并发度高的特点。基于某省农机作业数据对算法进行了有效性验证,结果表明,在数据规模达到 1×10^5 条、数据异常率为 5% 的情况下,算法异常识别率达到 0.94,且与已有清洗算法相比均方根误差更小。基于 Box - Behnken 方法设计试验,通过响应面分析得到回归模型,分析算法参数对均方根误差和运行时间的影响。基于二进制编码的混合遗传算法对参数进行优化,优化后的参数组合可使算法均方根误差达到 0.16、运行时间达到 0.13 s。该数据清洗方法能够为农机大数据平台的实时处理提供高质量数据支撑。

关键词: 农业机械; 大数据; 数据清洗

中图分类号: S24; TP391

文献标识码: A

文章编号: 1000-1298(2021)06-0035-08

OSID:



Experimental Optimization of Big Data Cleaning Method for Agricultural Machinery

YUAN Yanwei XU Ling JI Fuhua GUO Dafang AN Sa NIU Kang

(The State Key Laboratory of Soil - Plant - Machine System Technology,

Chinese Academy of Agricultural Mechanization Sciences, Beijing 100083, China)

Abstract: Data quality issues are the bottleneck hindering the development of agricultural machinery big data platforms. The existing data cleaning algorithms are not suitable for large-scale, multi-source heterogeneous, high-dimensional, and strong spatiotemporal correlation of agricultural machinery real-time streaming data. To this end, the source and characteristics of the abnormal data of agricultural machinery in complex environments were analyzed, the detection and correction technology of abnormal data was studied, and an online cleaning method for agricultural machinery operation data based on sliding window mechanism was proposed. The method determined abnormal data based on the principle of variance constraint; generated preliminary candidate data based on the principle of minimum change; based on the time correlation of data, the final repair value was obtained through AR and ARX model optimization; relying on the Flink distributed computing platform, it was suitable for large data throughput and high concurrency of agricultural machinery. The validity of the algorithm was verified based on the agricultural machinery operation data of a certain province. The results showed that when the amount of data reached 1×10^5 and the proportion of abnormal data was 5%, the abnormal recognition rate of the algorithm reached 0.94, and the root mean square error was smaller than that of the existing cleaning algorithm. The experiment was designed based on the Box - Behnken method, and the regression model was obtained through response surface analysis to study the influence of algorithm parameters on the root mean square error and time. The hybrid genetic algorithm based on binary coding optimized the parameters, and the optimized parameter combination can make the root mean square error of the algorithm reach 0.16 and the running time reach 0.13 s. The data cleaning method can provide high-quality data support for the real-time processing of the agricultural machinery big data platform.

Key words: agricultural machinery; big data; data cleaning

收稿日期: 2020 - 09 - 27 修回日期: 2020 - 10 - 28

基金项目: 国家重点研发计划项目(2017YFD0700205)

作者简介: 苑严伟(1980—),男,研究员,博士生导师,主要从事精准农业与农机测控技术研究,E-mail: yyw215@163.com

0 引言

农机大数据平台建设是推进现代农业生产信息化、智能化、精准化的重要环节^[1-7]。随着北斗系统、5G 通信、物联网等技术迅速发展,大数据平台不断完善,入网农机数量猛增,数据规模不足问题得到缓解,但数据质量问题成为阻碍平台发展的新瓶颈。据统计,大型数据集集中的错误率约为 5%^[8]。数据清洗能够减少“Garbage in, garbage out”现象,但所需时间很长,一般约占数据分析总时间的 60% ~ 80%^[9-10]。准确高效的数据清洗方法能够提高平台分析决策的可靠性和时效性,是农机大数据平台发展的重要基石。

国内外关于数据清洗的研究主要集中在异常数据检测^[11-17]方面,对异常数据修复的研究较少。处理异常数据的传统方式有直接保留、删除和人工填充等,其效果不够理想。ETL (Extract transform load) 工具虽然能够实现数据抽取、转换和加载,但是修复能力较弱,不能满足应用需求。文献[18]针对工业大数据中的高维时间序列数据,基于领域知识支持,开发了在线与离线相结合的数据清洗系统 Cleanits。文献[19]针对时间序列异常数据修复问题,提出将时间相关特性与最小变动原理相结合的 IMR (Iterative minimum repairing) 框架。文献[20]针对平滑过滤方法严重更改原始数据、且已有算法不支持流数据计算的问题,提出利用数据浮动速度函数修复高度异常数据的方法。文献[21-23]利用 Spark 分布式计算框架加快了大数据清洗流程。目前,已有数据清洗算法均未充分利用数据间的相互关系,对原始数据改动大,不适用于具有大规模、多源异构、高维度、强时空相关等特点的农机实时流数据。

为此,本文分析数据异常出现的主要场景及原因,根据农机作业数据的时间相关性和最小变动原则,提出一种数据清洗方法,依托 Flink 流计算平台实现数据的实时分析,通过试验验证算法的有效性,并对算法进行试验优化。

1 复杂作业条件下数据异常分析

由于应用领域存在差异,数据异常的概念缺乏统一定义。本文将数据异常定义为:在某一瞬时,服务器接收到的数据(或数据的某一部分)出现不完整、不准确、不合法等现象。

田间作业条件复杂,增加了数据异常发生的概率。在实际工作中,易发生数据异常的主要场景如下:

- (1)田间环境影响传感器检测精度。例如,地块周围存在高大树木或建筑物等,会遮挡一些传感器(如卫星定位传感器)的信号,导致数据出现跳变、离散或缺失。
 - (2)作业工况影响传感器检测精度。例如,作业过程中土壤或谷物产生的粉尘会干扰传感器(如光学传感器)的敏感元件;地面不平整或机器运转产生的振动也会影响传感器(如冲量式测产传感器)的检测精度,导致数据出现异常。该类型数据异常具体表现为数据出现零散、漂移或抖动。
 - (3)农机自身因素影响传感器检测精度。例如,机器的发动机或供电系统工作不稳定,引起电压波动,导致传感器(如电容式传感器)的检测性能随电压波动。该类型数据异常具体表现为数据出现漂移、丢失或抖动。
 - (4)田间环境影响信息上传质量。例如,田间网络信号差或存在电磁干扰,会影响数据传送的时效性和准确性,导致数据出现延时或缺失。
- 本文所讨论的数据异常不包括作业环境合理变化导致的数据波动。例如:在同一地块中,由于土壤肥力不同导致农作物产量变化,尽管数据看起来比较反常,但是并不属于异常数据。

2 基于滑动窗口的数据在线清洗算法

为实现数据在线清洗,提出基于滑动窗口实现的流数据异常识别和修复算法。针对农机作业数据以数值型为主的特点,基于方差约束原则识别异常数据;基于最小变动原则,对异常数据的原始值进行初步估算,生成候选数据;根据数据时间相关性,基于 AR (Autoregressive model)、ARX (Autoregressive model with exogenous input) 模型得到最优修复值。因此,算法分为识别异常数据、生成候选修正数据、数据迭代修正 3 个步骤,如图 1 所示。

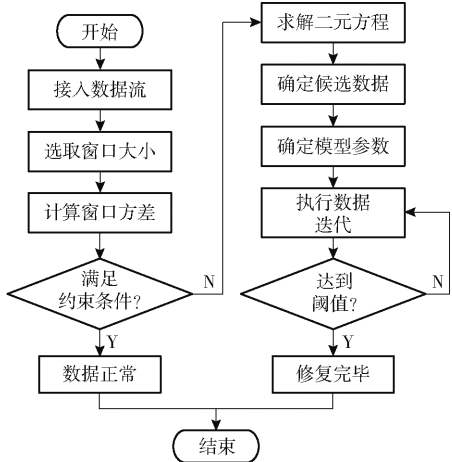


图 1 数据清洗流程

Fig. 1 Flow chart of data cleaning

2.1 异常数据动态识别

在每个数据对应的窗口区间内,进行方差检验,评估其是否为异常。通过窗口的滑动,可以依次评估每个数据,实现数据流的动态异常识别。

假设原始数据集 Γ 内前 w 个数据可靠,对于 Γ 内第 i ($i = w, w+1, w+2, \dots$) 个数据 γ_i ,选取大小为 w 的窗口 $D_i = \{d_{i1}, d_{i2}, \dots, d_{iw}\}$,其中 $d_{ij} = \gamma_{i-w+j}$,则 D_i 的方差为

$$\delta(D_i) = \frac{\sum_{j=1}^w (d_{ij} - \bar{D}_i)^2}{w} \quad (1)$$

式中 d_{ij} ——窗口 D_i 中第 j 个数据
 $\bar{D}_i$ ——窗口 D_i 中数据的平均值

依据以往所采集的同类数据的方差 v' ,选取方差阈值 v 作为评判基准, v 的取值应略大于 v' ,即 $v = \lambda v'$,其中 λ 为按经验选取的大于 1 的常数。当 $\delta(D_i) \leq v$ 时,则 γ_i 为正常数据;当 $\delta(D_i) > v$ 时,则 γ_i 为异常数据。

若 γ_i 为正常数据,则窗口向后滑动,对第 $i+1$ 个数据 γ_{i+1} 进行评估。若 γ_i 为异常数据,则修复该数据后,再向后滑动窗口。

2.2 异常数据候选修正集生成

若确定 γ_i 为异常数据,则设其为 x ,并令 $\delta(D_i) = v$,2 个原始解 x_1 和 x_2 求解公式为

$$\delta(D_i) = \frac{(d_{i1} - \bar{D}_i)^2 + \dots + (x - \bar{D}_i)^2 + \dots + (d_{iw} - \bar{D}_i)^2}{w} = v \quad (2)$$

假设 $x_1 \leq x_2$,对于异常数据 γ_i ,必然有 $\gamma_i < x_1$ 或 $\gamma_i > x_2$ 。当 $\gamma_i < x_1$ 时,根据最小变动原则,数据 γ_i 的候选值 γ'_i 取 x_1 。同理,当 $\gamma_i > x_2$ 时, γ'_i 取 x_2 。候选数据 γ'_i ,即

$$\gamma'_i = \begin{cases} x_1 & (\gamma_i < x_1) \\ x_2 & (\gamma_i > x_2) \end{cases} \quad (3)$$

通过此方法,依次将所有异常数据替换为候选数据,生成候选数据集 Γ' 。

2.3 候选修正集迭代优化

由于生成候选数据集的方法较为粗放,因此引入 AR 和 ARX 模型,利用异常数据 γ_i 前面 m 个数据,对 γ_i 的候选值 γ'_i 进行优化,保证数据修正准确可靠。

若 m 个数据均为正常数据,则使用 AR 模型更新候选数据 γ'_i ,得到最终修复值为

$$\gamma_i^* = C + \sum_{k=1}^m \phi_k \gamma'_{i-k} + \varepsilon_i \quad (4)$$

式中 γ_i^* ——最终修复值

C ——常量 m ——阶数

ϕ_k ——AR、ARX 模型参数

ε_i ——白噪声点

否则,使用 ARX 模型对第 $i-m$ 个至第 $i-1$ 个数据的原始值与候选值的差加权求和,更新候选数据 γ'_i ,得到最终修复值为

$$\gamma_i^* = \gamma'_i + \sum_{k=1}^m \phi_k (\gamma_{i-k} - \gamma'_{i-k}) + \varepsilon_i \quad (5)$$

其中 ϕ_k, m 可利用数学统计进行估算。样本数据集 Γ 的协方差函数符合 Yule-Walker 方程,可以得到

$$\begin{cases} \phi_1 \beta_0 + \phi_2 \beta_1 + \dots + \phi_p \beta_{p-1} = \beta_1 \\ \phi_1 \beta_1 + \phi_2 \beta_0 + \dots + \phi_p \beta_{p-2} = \beta_2 \\ \vdots \\ \phi_1 \beta_{p-1} + \phi_2 \beta_{p-2} + \dots + \phi_p \beta_0 = \beta_p \end{cases} \quad (6)$$

式中 $\beta_0, \beta_1, \dots, \beta_p$ —— Γ 的协方差函数
 转换为矩阵形式

$$\mathbf{A}_p \boldsymbol{\phi}(p) = \mathbf{B}_p \quad (7)$$

由于矩阵 $\mathbf{A}_p$ 对称且可逆,因此可得

$$\boldsymbol{\phi}(p) = \mathbf{A}_p^{-1} \mathbf{B}_p \quad (p = 1, 2, \dots) \quad (8)$$

此时,可以求出 ϕ_k ,其中 $\boldsymbol{\phi}(p)$ 的第 p 个分量 ϕ_{pp} ,即为偏相关函数。

根据 AR(m) 的特性可知,其偏自相函数 m 步截尾(在大于某个常数后快速趋于 0),因此可以将点(p, ϕ_{pp})在笛卡尔坐标系中标出。当存在某个 p 之后, ϕ_{pp} 无限接近 0,此时的 p 即为所求的阶数 m 。执行迭代,直至前后 2 次迭代的 γ'_i 小于阈值 τ 时,停止迭代。

数据异常清洗算法步骤如下:

(1) 输入 Γ, v, w, τ

(2) 处理 γ_i

$\gamma_i \rightarrow D_i$

if $\delta(D_i) \leq v$

$\gamma_i^* = \gamma'_{i,0} = \gamma_i, \gamma_i^* \in \Gamma'_0$

else

for l to $\dots$

$\delta(D_i) = v \rightarrow x_1, x_2 (x_1 \leq x_2)$

if $\gamma_i \leq x_1$

$\gamma'_{i,0} = x_1$

else

$\gamma'_{i,0} = x_2$

end

if $\gamma_{i-m}, \dots, \gamma_{i-1}$ do not have abnormal data

$\gamma'_{i,l} = C + \sum_{k=1}^m \phi_k \gamma'_{i-k} + \varepsilon_i$

else

$$\gamma_i^* = \gamma_i' + \sum_{k=1}^m \phi_k(\gamma_{i-k} - \gamma_{i-k}') + \varepsilon_i$$

if γ_i^*

$$\gamma_{i,l}^* \in \Gamma^*$$

break

end

end

end

end

end

(3)输出 Γ^*

3 试验与分析

3.1 试验材料

在现有农机大数据平台中,选取 2016—2019 年某省农机深松、保护性耕作等 8 种类型的作业数据,规模大于 1×10^9 条,代表性字段如表 1 所示。

表 1 农机数据基本信息

Tab.1 Basic information of agricultural machinery data		
作业类型	字段数	代表性字段
农机深松	12	经纬度、行驶/工作速度、面积
高效植保	13	经纬度、流量、面积、油耗
秸秆打捆	10	打捆速度、面积、经纬度、油耗
保护性种植	13	作业速度、面积、经纬度、时间
马铃薯播种	13	播种量、经纬度、面积、油耗
马铃薯收获	13	收获速度、产量、经纬度、油耗
小麦收获	11	收获速度、产量、经纬度、油耗
籽粒直收	14	收获速度、产量、经纬度、油耗

为适应农机作业数据吞吐量大、并发度高的特点,将算法迁移至大数据流计算平台 Flink 上,依托 Flink 集群的分布式特性,保证算法快速准确执行。传感器通过 TCP/IP 协议将海量数据传输至 Kafka 集群进行分组管理。然后,Flink 消费者集群接收数据并运行算法,实现流数据清洗。系统部署在阿里云服务器上,相关配置如表 2 所示。

表 2 试验环境

Tab.2 Experimental environment	
服务器配置参数	型号/数值
操作系统	CentOS 7.0
CPU	Intel 酷睿 i7 8700K
主频/GHz	3.7
内存/GB	32(Flink 集群 256 GB)
硬盘/TB	20(Flink 集群 10 TB)
Flink 组件	1.4.0
Kafka 组件	2.11 - 1.0.0
ZooKeeper 组件	3.4.6

3.2 算法有效性验证

3.2.1 验证方法

算法有效性包括异常数据识别有效性和修复有效性。使用精确率 P_1 、召回率 R 和综合性指标 F_1 评价异常数据识别有效性。使用均方根误差 E_{RMSE} 评价数据修复有效性,计算式为

$$E_{\text{RMSE}} = \sqrt{\frac{\sum_{i=1}^n (\gamma_i - \gamma_i^*)^2}{n}}$$

(9)

式中 n ——异常数据个数
 γ_i^* ——修复后的异常数据

算法有效性验证方案流程如图 2 所示。首先,选取一定规模的正常数据,并人工对其进行预处理,将一定比例的正常数据修改至异常,同时标记正常数据为 1,异常数据为 0。将预处理后的数据作为试验集,使用算法完成数据清洗。进行多次清洗后,取各评价指标的平均值进行分析。

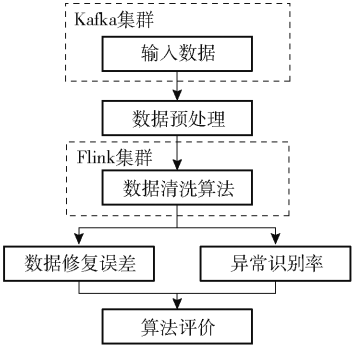


图 2 算法有效性验证方案
Fig.2 Algorithm verification scheme

为更加具体地展现本算法的修复效果,采用基于平滑的清洗算法 SWAB^[24] 和基于否定约束的全局清洗算法 Holistic^[25] 与本算法进行横向对比。

3.2.2 验证结果

(1)选取不同规模的试验集,在数据预处理时将 5% 的数据修改至异常,取窗口大小为 100,阶数取值为 4,阈值取值为 0.1。在每一规模水平下进行 3 次试验,分别对 P_1 、 R 、 F_1 和 E_{RMSE} 取平均值。

如图 3 所示, P_1 、 F_1 随数据量增加而增大。当数据规模达到 1×10^5 条, P_1 在 0.94 附近趋于稳定; R 一直处于 0.9~0.95 区间内; F_1 大于 0.92。说明算法在大规模数据集中具有较高的异常识别率。

均方根误差随数据规模的变化如图 4 所示。随着数据规模的增大,3 种算法的均方根误差均减小,且在数据达到一定规模后,均方根误差的变化趋缓。SWAB 算法均方根误差始终较高。在数据规模小于 1×10^5 条时,Holistic 与本文算法表现相近,但在数据规模大于等于 1×10^5 条时,后者的均方根误差明

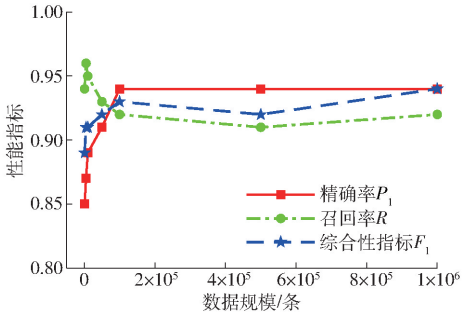


图3 异常数据识别指标

Fig. 3 Abnormal data identification index

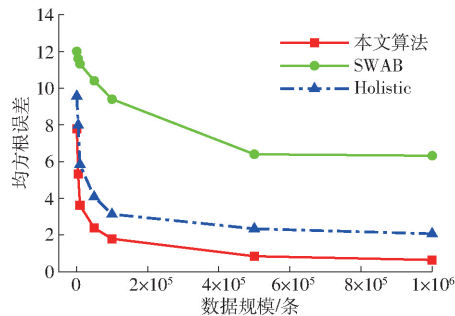


图4 均方根误差与数据规模的关系曲线

Fig. 4 Relationship between root mean square error and data size

显更小,表明本文算法修复效果较好。

(2)选取规模为 1×10^5 条的试验集,在窗口大小为 100、阶数为 4、阈值为 0.1 时,选取不同数据异常率进行试验。试验结果如图 5 所示,同一数据规模下,3 种算法的均方根误差均随数据异常率的增大而增大,SWAB 算法最高,Holistic 算法次高,本文算法始终最低,说明所提出的算法在修复准确性方面有效,且数据异常率越低表现越好。

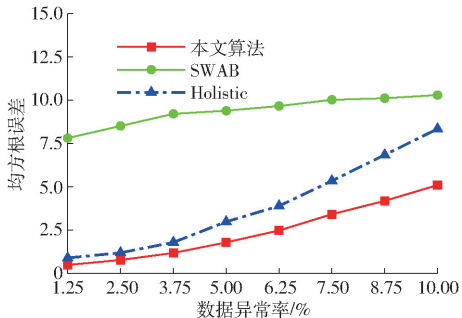


图5 数据异常率与均方根误差的关系曲线

Fig. 5 Relationship between data anomaly rate and root mean square error

3.3 算法试验优化

由前述分析,数据规模处于较高水平、错误率处于较低水平时,二者对均方根误差的影响很小,并且实际情况中数据规模与错误率不可控。因此,对算法的优化主要考虑阶数、阈值、窗口大小对均方根误差 E_{RMSE} 和算法运行时间 T 的影响。采用 Box - Behnken 原理设计试验,各试验因素编码如表 3 所

示,试验设计与结果如表 4 所示, A 、 B 、 C 为因素 m 、 τ 、 w 编码值。

表 3 因素编码

Tab. 3 Coding of factors

编码	因素		
	阶数 m	阈值 τ	窗口大小 w
0	0	0	0
0.5	4.5	0.125	2 000
1	9.0	0.250	4 000

表 4 试验设计与结果

Tab. 4 Experiment design and result

试验号	因素			E_{RMSE}	T/s
	A	B	C		
1	1	0.5	1	1.45	15.04
2	0.5	1	0	11.69	2.10
3	0.5	0.5	0.5	1.36	2.56
4	0	1	0.5	20.39	2.36
5	0.5	0.5	0.5	1.57	1.90
6	0.5	0.5	0.5	4.67	3.53
7	1	1	0.5	15.60	16.03
8	0.5	0.5	0.5	1.57	12.53
9	0	0.5	1	9.94	1.79
10	0.5	0	1	2.18	14.00
11	0.5	0.5	0.5	1.16	1.89
12	0.5	1	1	11.76	2.71
13	0	0.5	0	10.16	4.81
14	0	0	0.5	12.03	18.62
15	1	0.5	0	3.88	17.19
16	0.5	0	0	6.10	18.39
17	1	0	0.5	0.86	30.86

对试验结果进行分析,选择回归模型进行拟合。均方根误差和算法运行时间的方差分析结果如表 5、6 所示,所建立的模型均显著 ($P \leq 0.05$),且失

表 5 均方根误差 E_{RMSE} 方差分析

Tab. 5 Variance analysis of root mean square error

来源	平方和	自由度	均方	F	P
模型	557.44	9	61.94	25.81	0.000 1
A	118.15	1	118.15	49.23	0.000 2
B	183.10	1	183.10	76.30	<0.000 1
C	5.28	1	5.28	2.20	0.181 4
AB	10.19	1	10.19	4.25	0.078 3
AC	1.23	1	1.23	0.51	0.497 6
BC	4.00	1	4.00	1.67	0.237 7
A^2	77.50	1	77.50	32.30	0.000 7
B^2	144.83	1	144.83	60.35	0.000 1
C^2	0	1	0	0	0.997 2
残差	16.80	7	2.40		
失拟	8.22	3	2.74	1.28	0.395 5
误差	8.58	4	2.14		
总和	574.24	16			

表6 算法运行时间 T 方差分析

Tab.6 Variance analysis of algorithm running time					
来源	平方和	自由度	均方	F	P
模型	1 127. 24	9	125. 25	10. 36	0. 002 8
A	332. 01	1	332. 01	27. 47	0. 001 2
B	430. 41	1	430. 41	35. 62	0. 000 6
C	10. 02	1	10. 02	0. 829 2	0. 392 8
AB	0. 507 4	1	0. 507 4	0. 042	0. 843 5
AC	0. 191 7	1	0. 191 7	0. 015 9	0. 903 3
BC	6. 27	1	6. 27	0. 518 4	0. 494 8
A^2	174. 88	1	174. 88	14. 47	0. 006 7
B^2	153. 57	1	153. 57	12. 71	0. 009 2
C^2	6. 28	1	6. 28	0. 519 3	0. 494 5
残差	84. 59	7	12. 08		
失拟	1. 78	3	0. 593 2	0. 028 7	0. 992 5
误差	82. 81	4	20. 7		
总和	1 211. 83	16			

拟项不显著,证明模型所拟合的回归方程与实际相符,能准确反映均方根误差、时间与阶数、阈值、窗口大小之间的关系。在保证模型可靠的前提下,为使回归模型更好地对试验结果进行预测,剔除不显著项,对回归模型做优化调整,分别得到 E_{RMSE} 和 T 的回归模型为

$$E_{\text{RMSE}} = 11.28 - 24.85A - 13.89B + 17.16A^2 + 23.46B^2$$

$(A,B \in [0,1])$

(10)

$$T = 17.22 - 12.64A - 38.57B + 25.52A^2 + 23.90B^2$$

$(A,B \in [0,1])$

(11)

将回归模型映射回原空间可得

$$E_{\text{RMSE}} = 11.28 - 2.76m - 55.56\tau + 0.21m^2 + 375.36\tau^2$$

$(m \in [0,9], \tau \in [0,0.25])$

(12)

$$T = 17.22 - 1.4m - 154.28\tau + 0.32m^2 + 382.4\tau^2$$

$(m \in [0,9], \tau \in [0,0.25])$

(13)

由图6可以看出,当阈值 τ 一定时,随着阶数 m 的增大,均方根误差先减少后增大。这说明过大的阶数并不会持续增加算法的准确度。当阶数一定时,均方根误差随阈值增大而增大,这说明选取较小的阈值能提高算法准确性。

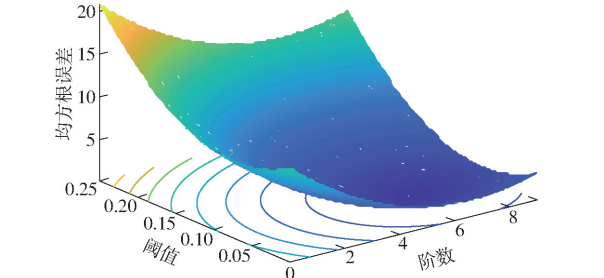


图6 阶数与阈值对均方根误差的影响
Fig.6 Influence of order and threshold on root mean square error

由图7可以看出,当阶数一定时,随着阈值逐渐变大,算法运行时间逐渐减少;当阈值较小时,需要较高的时间成本。当阈值一定时,算法运行时间随阶数增大而增大,结合阶数对均方根误差的影响,说明选取过大的阶数导致算法准确性差且效率低。

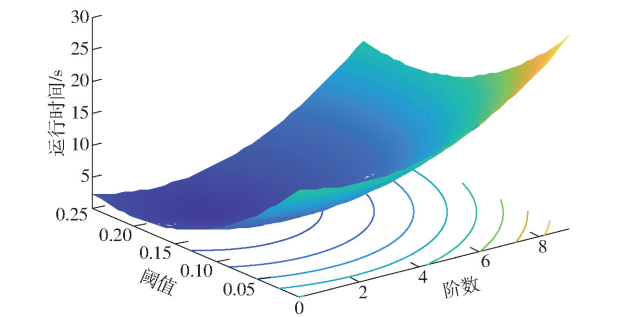


图7 阶数与阈值对运行时间的影响
Fig.7 Influence of order and threshold on running time

云服务平台使用目的不同,所要求的性能指标不同,有些注重数据准确性,有些更关注实时性。为此,本文采用二进制编码的混合遗传算法,分别对式(12)、(13)所示的模型进行优化求解,确定不同性能指标下的参数组合,优化过程如图8、9所示。设定种群个体数目为15,交叉概率为0.8,变异概率

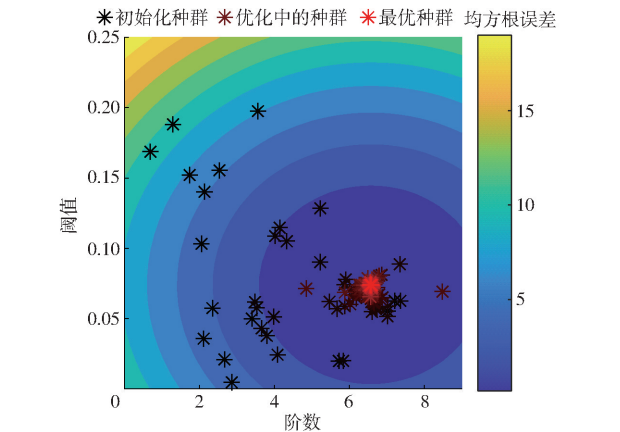


图8 均方根误差迭代优化
Fig.8 Iterative optimization of root mean square error

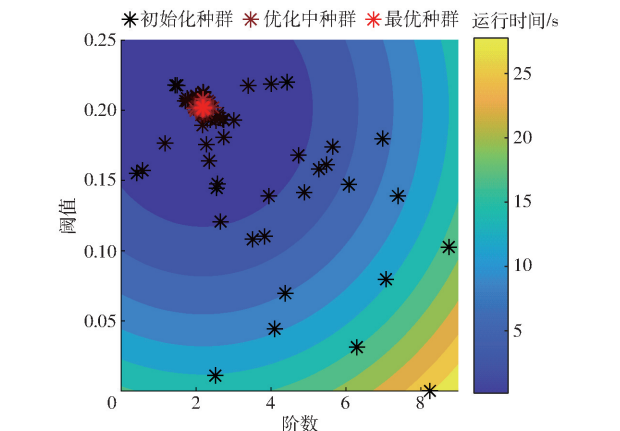


图9 算法运行时间迭代优化
Fig.9 Iterative optimization of algorithm runtime

为0.08。迭代求解得到,当阶数为6.6,阈值为0.07时,均方根误差最小为0.16;当阶数为2,阈值为0.2,算法运行时间最小为0.13 s。

4 结论

(1)研究异常数据检测及修正技术,提出一种基于滑动窗口机制的数据在线清洗算法,并依托Flink分布式计算平台,加速数据的实时清洗,以适应农机大数据高并发、吞吐量大的特点。

(2)试验表明,针对规模为 1×10^5 条、异常率为5%的数据集,算法窗口大小取100、阶数取4、阈值取0.1时,精确率、召回率和综合性指标均满足数据清洗要求。与SWAB算法和Holistic算法修复后的

均方根误差的对比表明,本文算法的均方根误差更小,从而证明了本文算法的有效性。本文算法的均方根误差随数据规模的增大而减小,随数据异常率的增大而增大,说明该算法适用于异常率较低的大规模数据集。

(3)基于Box-Behnken原理设计试验,分别建立均方根误差、算法运行时间与阶数、阈值、窗口大小之间的响应曲面回归模型。利用基于二进制编码的混合遗传算法求出满足各性能指标的多参数最优组合,当阶数为6.6、阈值为0.07时,均方根误差最小为0.16,当阶数为2、阈值为0.2时,算法运行时间最短为0.13 s。

参 考 文 献

- [1] HAOTIAN Y, SHUMING X, AKWASI F S, et al. A consortium blockchain-based agricultural machinery scheduling system[J]. Pubmed, 2020, 20(9):1-3.
- [2] SUN F, WANG X, ZHANG R. Task scheduling system for UAV operations in agricultural plant protection environment[J]. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(5): 1-15.
- [3] 曹健. 黑龙江省农机管理调度指挥平台[N]. 中国农机化导报, 2019-09-09(006).
- [4] 程亚维, 王东霞. 基于Java的农机作业调度管理Web平台架构技术研究[J]. 农机化研究, 2019, 41(10): 224-228. CHENG Yawei, WANG Dongxia. Research on Web platform architecture of agricultural machinery scheduling management based on Java[J]. Journal of Agricultural Mechanization Research, 2019, 41(10): 224-228. (in Chinese)
- [5] 张璠, 李蔓, 常淑惠, 等. 农机作业路径规划策略研究——基于智慧农机大数据平台[J]. 农机化研究, 2020, 42(12): 17-22. ZHANG Fan, LI Man, CHANG Shuhui, et al. Research on route planning strategy of agricultural machinery operation—based on intelligent agricultural machinery big data platform[J]. Journal of Agricultural Mechanization Research, 2020, 42(12): 17-22. (in Chinese)
- [6] 张昱婷, 滕桂法. 基于大数据的农机作业管理与决策系统[J]. 河北农机, 2018(7): 13-14. ZHANG Yuting, TENG Guifa. Agricultural machinery operation management and decision-making system based on big data[J]. Hebei Agricultural Machinery, 2018(7): 13-14. (in Chinese)
- [7] 周倩, 周林妹, 康丽军. 基于大数据和QT的农机配件供应链云平台系统设计[J]. 农机化研究, 2019, 41(7): 207-211. ZHOU Qian, ZHOU Linmei, KANG Lijun. Design of cloud platform system for supply chain of agricultural machinery parts based on large data and QT[J]. Journal of Agricultural Mechanization Research, 2019, 41(7): 207-211. (in Chinese)
- [8] 邹旺. 基于聚类的数据清洗算法的研究[D]. 哈尔滨: 哈尔滨工程大学, 2015. ZOU Wang. Research on data cleaning algorithm based on clustering[D]. Harbin: Harbin Engineering University, 2015. (in Chinese)
- [9] BENNANE A, YACOUT S. Processing missing and inaccurate data in a condition based maintenance database[C]//The 40th International Conference on Computers & Industrial Engineering. IEEE, 2010:1-5.
- [10] 冯宪凯, 黄树成. 基于DBSCAN的缺失值填充算法研究[J]. 计算机与数字工程, 2020, 48(7): 1572-1575, 1686. FENG Xiankai, HUANG Shucheng. Research on missing value filling algorithm based on DBSCAN[J]. Computer and Digital Engineering, 2020, 48(7): 1572-1575, 1686. (in Chinese)
- [11] O'REILLY C, GLUHAK A, IMRAN M A, et al. Anomaly detection in wireless sensor networks in a non-stationary environment[J]. IEEE Communications Surveys & Tutorials, 2014, 16(3): 1413-1432.
- [12] SHAHID N, NAQVI I H, QAISAR S B. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey[J]. Artificial Intelligence Review, 2015, 43(2): 193-228.
- [13] 吴金娥, 王若愚, 段倩倩, 等. 基于反向 k 近邻过滤异常的群数据异常检测[J]. 上海交通大学学报, 2021, 55(5): 598-606. WU Jin'e, WANG Ruoyu, DUAN Qianqian, et al. Collective data anomaly detection based on reverse k -nearest neighbor filtering[J]. Journal of Shanghai Jiaotong University, 2021, 55(5): 598-606. (in Chinese)
- [14] LIU L. A new method for outlier detection on time series data[D]. Vancouver: Simon Fraser University, 2015.
- [15] 施晓斌, 吴丹萍, 程红举. 无线传感器网络中基于模式频繁度的异常检测方法[J]. 信息技术与网络安全, 2018, 37(1): 70-75. SHI Xiaobin, WU Danping, CHENG Hongju. Anomaly detection method based on pattern frequencies in wireless sensor

- networks[J]. Information Technology and Network Security, 2018, 37(1): 70–75. (in Chinese)
- [16] ISLAM R U, HOSSAIN M S, ANDERSSON K. A novel anomaly detection algorithm for sensor data under uncertainty[J]. Soft Computing, 2018, 22(5): 1623–1639.
- [17] YUAN H, ZHAO X, YU L. A distributed Bayesian algorithm for data fault detection in wireless sensor networks[C]//2015 International Conference on Information Networking (ICOIN). IEEE, 2015: 63–68.
- [18] DING X, WANG H, SU J, et al. Cleanits: a data cleaning system for industrial time series[J]. Proceedings of the VLDB Endowment, 2019, 12(12): 1786–1789.
- [19] ZHANG A, SONG S, WANG J, et al. Time series data cleaning: from anomaly detection to anomaly repairing[J]. Proceedings of the VLDB Endowment, 2017, 10(10): 1046–1057.
- [20] SONG S, ZHANG A, WANG J, et al. Screen: stream data cleaning under speed constraints[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015: 827–841.
- [21] 金翰伟. 基于 Spark 的大数据清洗框架设计与实现[D]. 杭州: 浙江大学, 2016.
- JIN Hanwei. Bigdata cleaning framework design and implementation based on Spark[D]. Hangzhou: Zhejiang University, 2016. (in Chinese)
- [22] 李志鹏. 基于 Spark 的可定义规则数据清洗系统的设计与实现[D]. 成都: 电子科技大学, 2019.
- LI Zhipeng. Design and implementation of a data cleaning system with definable rules based on Spark[D]. Chengdu: University of Electronic Science and Technology of China, 2019. (in Chinese)
- [23] QU Z Y, WANG Y W, WANG C, et al. A data cleaning model for electric power big data based on Spark framework[J]. International Journal of Database Theory and Application, 2016, 9(3): 137–150.
- [24] KEOGH E, CHU S, HART D, et al. An online algorithm for segmenting time series [C]//Proceedings 2001 IEEE International Conference on Data Mining. IEEE, 2001: 289–296.
- [25] CHU X, ILYAS I F, PAPOTTI P. Holistic data cleaning: putting violations into context[C]//2013 IEEE 29th International Conference on Data Engineering (ICDE). IEEE, 2013: 458–469.

(上接第 25 页)

- [28] MIŠEKIS J, GLETTE K, ELLE O J, et al. Multi 3D camera mapping for predictive and reflexive robot manipulator trajectory estimation[C]//IEEE Symposium Series on Computational Intelligence, 2016: 1–8.
- [29] ALENY G, FOIX S, TORRAS C. ToF cameras for active vision in robotics[J]. Sensors and Actuators A: Physical, 2014, 218: 10–22.
- [30] CHEN X, CHAUDHARY K, TANAKA Y, et al. Reasoning-based vision recognition for agricultural humanoid robot toward tomato harvesting[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2015: 6487–6494.
- [31] LEHNERT C, ENGLISH A, MCCOOL C, et al. Autonomous sweet pepper harvesting for protected cropping systems[J]. IEEE Robotics & Automation Letters, 2017, 2(2): 872–879.
- [32] SA I, LEHNERT C, ENGLISH A, et al. Peduncle detection of sweet pepper for autonomous crop harvesting-combined colour and 3D information[J]. IEEE Robotics & Automation Letters, 2017, 2(2): 765–772.
- [33] LEHNERT C, MCCOOL C, PEREZ T. Lessons learnt from field trials of a robotic sweet pepper harvester [J]. arXiv preprint arXiv:1706.06203, 2017.
- [34] SORIA P R, SUKKAR F, MARTENS W, et al. Multi-view probabilistic segmentation of pome fruit with a low-cost RGB–D camera[C]//Proceedings of the Iberian Robotics Conference, 2017: 320–331.
- [35] 刘继展, 朱新新, 袁妍. 枝上柑橘果实深度球截线识别方法[J/OL]. 农业机械学报, 2017, 48(10): 32–39.
- LIU Jizhan, ZHU Xinxin, YUAN Yan. Depth-sphere transversal method for on-branch citrus fruit recognition [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(10): 32–39. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20171004&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2017.10.004. (in Chinese)
- [36] LIU J Z, YUAN Y. Experiments and analysis of close-shot identification of on-branch citrus fruit with RealSense[J]. Sensors, 2018, 18(5): 1510.