基于 BP 神经网络的录井异常数据检测方法研究

李春生,邹林浩,张可佳,高雅田,刘 涛,豆立宪
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:在石油钻井工程中,由于技术和设备的客观因素,导致录井数据经常出现异常值,影响了录井解释评价精度。针对该问题,提出了一种基于 BP 神经网络的录井异常数据处理方法。为了在构建数据环节中提供准确且可信的工程数据,研究了录井异常数据的产生原因及异常数据的表征,并且通过对比格鲁布斯法、K-means 聚类算法以及 BP 神经网络等方法的特点,选择 BP 神经网络作为异常值处理的方法。通过模型预测的录井数据误差平方值与样本数据的均方根误差进行比较,来确定数据的异常情况,保证检测异常点的合理性。经实验验证和同类算法的比较,表明了 BP 神经网络模型可以实现检测录井异常点数据,且检测异常点的准确率高于同类算法,处理异常点结果可信,能够有效解决因异常点数据所带来的问题。

关键词:异常点检测;录井工程数据;BP 神经网络;格鲁布斯法;K-means 聚类算法
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)06-0173-06
doi:10.3969/j.issn.1673-629X.2022.06.029

Research on Detection Method of Logging Anomaly Data Based on BP Neural Network

LI Chun-sheng,ZOU Lin-hao,ZHANG Ke-jia,GAO Ya-tian,LIU Tao,DOU Li-xian
(School of Computer and Information Technology,Northeast Petroleum University,Daqing 163318,China)

Abstract:In the oil drilling engineering,because of the objective factors of technology and equipment,abnormal values often appear in the logging data,which affects the accuracy of logging interpretation and evaluation. Aiming at this problem,a method of logging anomaly data processing based on BP neural network is proposed. In order to provide accurate and reliable engineering data in the construction of data,we study the causes of logging abnormal data and the characterization of abnormal data,and select BP neural network as the method of outlier processing by comparing the characteristics of Grubbs method,K-means clustering algorithm,BP neural network and other methods. By comparing the square error of the logging data predicted by the model with the root mean square error of the sample data,the abnormal situation of the data can be determined to ensure the rationality of the abnormal points detected. The experimental verification and comparison with the similar algorithms show that the BP neural network model can detect logging anomaly data,and the accuracy of detecting anomaly points is higher than that of the similar algorithms. The results of handling anomaly points are reliable,and it can effectively solve the problems caused by the abnormal point data.

Key words:anomaly detection;logging engineering data;BP neural network;Grubbs method;K-means clustering algorithm

0 引 言

随着信息技术的不断提升,在生活和生产过程都产生大量的数据,在大量的数据中往往存在异常数据,这些异常数据都或多或少地影响着人们的生产及生活[1]。异常点检测能够在大量的数据中快速、准确地找出异常数据并进行处理,解决异常数据带来的问题或事故。

目前在录井物性评价方法中,机械比能比值方法和功交汇方法是评价所依据的定量化标准,但是工程数据中有大量异常点数据,无法判断其是否准确,导致在计算物性评价时产生了不理想的结果,甚至对岩性识别、物性评价、油气性监测等工作也产生了不同程度的影响,因此越来越多的录井科研人员开始研究处理异常值问题。当前处理录井异常数据的方法主要是利

收稿日期:2021-07-13 **修回日期**:2021-11-18
基金项目:国家自然科学基金项目(51774090);黑龙江省青年创新人才培养计划(UNPYSCT-2020144);黑龙江省省属本科高校基本科研业务费东北石油大学引导性创新基金(2021YDL-12)
作者简介:李春生(1960-),男,博士,教授,博导,研究方向为多智能体系统方法论、模式挖掘技术等研究、软件集成技术;邹林浩(1998-),男,研究生,CCF 会员(H7074G),研究方向为大数据与智慧城市、数据挖掘。

用 3σ 准则和人工处理方法。3σ 准则的处理原理是把均值±(3 * 标准差)范围外的数据作为异常值,并计算出相邻 5 个数据点的平均值作为替代值,该方法误差较大,且经常存在异常值漏检。还有部分录井解释人员使用人工处理方法,解释人员将录井数据以折线图的方式展现出来,并观察找出偏差较大的异常值剔除,此方法效率低、成本高,极大地耗费了人力物力,因此目前处理录井异常数据的方法不完全合适,需要采用科学的方法对这些工程数据进行异常点处理。

针对"如何更快速、精准地处理录井数据中异常数据"问题,该文提出了一种基于 BP 神经网络的异常值处理方法。通过建立 BP 神经网络模型,并将 BP 神经网络数据样本的均方根误差作为阈值判断数据的异常情况。实验表明,基于 BP 神经网络方法处理录井数据异常值精准度更高、误差更小,论证了该方法的有效性和应用价值。

1 相关工作

由于各种异常值检测方法针对的测定值都不同,效果也各不相同,所以在测定值和异常值检测方法不同的情况下得到的结论和效果也是不同的,很难找出最优的异常值检测方法。根据录井中异常数据的特点,本节主要根据异常数据的表征比较不同异常点检测方法,选择出可靠的检测异常值数据方法来处理录井工程数据。

1.1 异常数据产生原因及表征分析

在石油钻井过程中,能导致录井工程数据出现异常值的原因有很多,其中包括钻具的疲劳可能使扭矩数值增大;钻具的折断可能使扭矩减小,同时钻速增大;井壁垮塌、井漏可能导致转速加快、扭矩减小,同时钻压降低等现象。

录井异常数据的表征主要分为突变值数据异常和整体偏差异常两种情况:(1)突变值数据异常,指的是录井数据在某一井深突然出现较大程度的升高或降低,但是在很短的时间内又恢复到正常数值的情况。比如操作不当导致工具掉落等情况,可能使录井数据产生突变值异常数据。(2)整体偏差异常,指录井数据普遍整体的升高或降低。比如钻具的折断、卡钻等情况,但是当地层发生改变时,也有可能导致录井数据的整体改变,所以在判断数据异常时应该具体情况具体分析。

1.2 检测异常点方法分析

(1)格鲁布斯检验法。

格鲁布斯法[2],是建立在两方差比式 $S_n^2/S_{(n-1)}^2$ 的基础之上的。

$$\frac{S_n^2}{S_{(n-1)}^2}=\frac{n-2}{n-1}/\left[1-\frac{n}{(n-1)^2}G^2\right] \tag{1}$$

$$G=\frac{x_d-\overline{x_n}}{S_n} \tag{2}$$

式中,S_n^2 是用 n 个测定值计算的方差,$S_{(n-1)}^2$ 是用 $(n-1)$ 个测定值计算的方差,G 是格鲁布斯检验法的检验统计量。由式(1)与式(2)知道,用 G 与用式 $S_n^2/S_{(n-1)}^2$ 作统计量进行检验是等效的。因为 S_n^2 与 $S_{(n-1)}^2$ 两者都是 σ^2 的,其比值应≒1。若一组数据中有一个以上的异常值,方差 $S_{(n-1)}^2$ 中至少包括了一个以上的异常值在内,使 $S_{(n-1)}^2$ 变大,但是 $S_n^2/S_{(n-1)}^2$ 的比值不一定变大,于是会使得数据中的一些异常值检验不出来。所以,当一组数据中有一个以上的异常值时,格鲁布斯检测法就不会是最有效的异常值检测方法。

(2)K-均值聚类。

K-means 聚类算法是由 J. B. MacQueen 在 1967 年提出的[3]。K-means 算法的中心思想是将指定的数据集合按照规定分为 K 组,数据集 M 由 X 个 a 维数据所构成。每一个类都被当作一个分组,其中用 C_k 来表示每一个类,也叫做簇,并且每个类 C_k 都有一个中心 O_i [4-5]。其中 m_i-O_k 为两个数据点的欧氏距离,计算类中数据点到聚类中心 O_i 的误差平方和,计算公式如下所示:

$$J(C_k)=\sum_{m_i\in C_i} m_i-O_k^2 \tag{3}$$

(3)BP 神经网络。

BP 神经网络算法:人工神经网络(artificial neural network)在能源应用、图像分析、医学工程等领域中有着广泛的应用[6-8]。它是一个由大量节点(神经元)构成的非线性系统,具有强大的容错性,并且具有自学习、自组织、自适应的能力[9]。神经网络通过网络学习得到结果的误差,反向传播到隐含层,通过反复的训练和改变权值、阈值,最终确定了最小误差对应的系数,完成对目标的学习。当 BP 神经网络的层数和神经元数较多时,会导致程序的循环嵌套,但是可以利用 MATLAB 中自带的工具箱,使建立 BP 神经网络模型变得简单便捷。因此,可直接通过 MATLAB 完成对 BP 神经网络的训练和学习。其伪代码如算法 1 所示:

算法 1:BP 神经网络(误差反向传播算法)。

input:训练集 $D=\{(x_k,y_k)\}_{k=1}^m$;学习率 η

output:多层前馈神经网络

1 begin

2 在(0,1)范围内随机初始化网络中的所有连接权值的阈值;

3 repeat

4 for all $(x_k,y_k)\in D$ do

5 计算当前样本输出;

6 计算输出神经元的梯度;
7 计算隐层的神经元梯度项;
8 更新权值;
9 更新阈值;
10　end for
11　until 达到停止条件。
12 end

1.3　数据异常点检测处理方法选择

综上所述,当在一组测定值中存在一个以上的异常值时,格鲁布斯检验法检测异常值不是高效的[2]。K-means 聚类算法最大的问题就是 K 值的选择问题,当某段数据整体提升或者下降时,也就无法判别 K 的取值,严重影响查找异常点的准确率。

在过去的一些文献中还介绍了一些其他的检验异常值的方法如 $2.5d$ 和 $4d$ 法[10]、拉依达法[11]、肖维特(Chauvenet)法[12]等。已知算术平均差 d 与标准差 σ 的关系是 $\sigma = 1.25d$,$2.5d$ 和 $4d$ 也就等于 2 倍的方差和 3.2 倍的方差,其概率分别为 95% 与 99.9%。故 $2.5d$ 和 $4d$ 法也就相当于 2σ 与 3.2σ 法,$2.5d$ 和 $4d$ 法只适用于检测较大数据量测定值中的异常值。拉依达检验法检测异常值的基本准则是 $|x_d - \bar{x}| > 3s$,当检测次数较少时,即使在测定值中有异常值,也有可能无法检测出来,容易发生错判和漏判的情况,它只适用于检测次数较多的情况。

针对以上这些问题,利用 BP 神经网络算法可以很好地解决[13],异常点检测实际上可以看作对未知数据的预测,计算出阈值进行比较,关键在于建立检测异常点数据的模型,根据建立好的神经网络模型就可以准确地判断数据异常情况。

2　数据异常检测建模

2.1　BP 神经网络异常点检测模型

(1)样本的确定。

选取合适的训练样本,为解决录井数据中不同数据存在数量级不同的问题,应该首先对录井数据进行归一化,保证所有数据都在 0-1 范围内。

$$[p_n, \min(p), \max(p), t_n, \min(t), \max(t)] = \text{premnmx}(p, t) \tag{4}$$

式中,p 是输入数据;t 是目标输出数据;t_n 是 t 归一化后的数据;p_n 是 p 归一化后的数据。

(2)网络层数的确定。

根据 Kolmogorov 定理可知,3 层的 BP 神经网络可以实现任意非线性映射,所以先确定网络层数为 3 层[14]。

$$\text{net} = \text{newff}(\text{minmax}(p_n), [n, 1], \{\text{'tansig'}, \text{'purelin'}\}, \text{'trainlm'}) \tag{5}$$

式中,net 为网络建立函数;n 为隐含层神经元数目;tansig 为正切 S 型传递函数;trainlm 为基于 L-M 规则的训练前向网络函数;purelin 为线性传递函数;1 为隐含层个数。

(3)隐含层神经元数的确定。

隐含层的神经元数直接影响着网络的容量、泛化能力、学习速度以及输出性能等。参考公式[15]:

$$n = \sqrt{n_i + n_o} + a \tag{6}$$

式中,a 为 1 ~ 10 之间的常数;n_o 为输出层神经元数;n_i 为输入层神经元数;n 为隐含层神经元数。本模型将工程异常数据作为输入,将工程标准数据作为输出,则 n_i 为 4,n_o 为 4,n 为 8,初步建立 4-8-4 的 BP 神经网络模型,如图 1 所示。

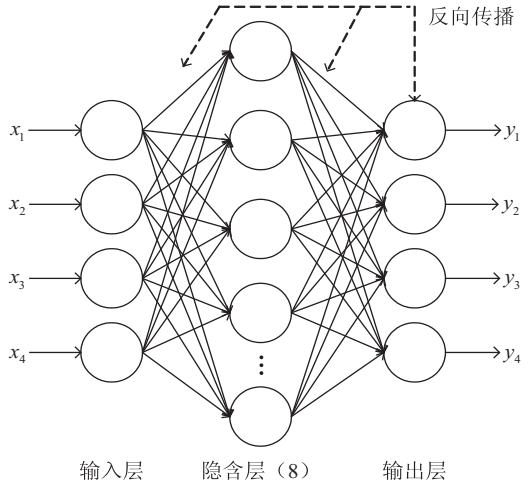

图 1　BP 神经网络预测模型网络

(4)其他网络训练指标的确定。

本模型 2 次显示之间的训练次数确定为 50,学习速率确定为 0.005,μ 的初始值确定为 0.9,网络目标性能确定为 0.000 02,训练代数确定为 4 000。在训练过程中,为提高效果可对其各项指标进行适当修改。

(5)训练网络。

本模型只要调用 TRAINGDM 算法训练 BP 网络。

$$[\text{netTR}] = \text{train}(\text{net}, p_n, t_n) \tag{7}$$

式中,net 为网络建立函数;p_n 是 p 归一化后的数据;t_n 是 t 归一化后的数据。

因为输入的训练数据是归一化处理后的数据,所以在模型训练后得出的输出数据应该进行数据还原。

$$A_n = \sin(\text{net}, p_n, t_n) \tag{8}$$

$$A = \text{postmnmx}(A_n, \min(t), \max(t)) \tag{9}$$

式中,A 为 A_n 反归一化后的输出数据,其与目标输出数据 t 之间的拟合程度由拟合度 R 指标描述;A_n 为 p_n 的网络仿真结果。

(6)检验网络。

检验网络是测试训练网络学习的结果,输入数据

进行网络模型测验,完成对输出数据的预测[16]。

$$p_{2n} = \text{tramnmx}(p_2, \text{minp}, \text{maxp}) \tag{10}$$

$$A_{2n} = \text{sim}(\text{net}, p_{2n}) \tag{11}$$

$$A_2 = \text{postmnmx}(A_{2n}, \min(t), \max(t)) \tag{12}$$

式中,p_2 为检验样本或预测样本的输入数据;p_{2n} 为 p_2 归一化后的数据;A_{2n} 为网络对 p_{2n} 的仿真结果;A_2 为输出数据;其他变量意义同前。若为检验网络,则将 A_2 与检验样本的目标输出数据 t_2 进行比较,可获得检验误差;若为预测网络,则 A_2 为预测值。

2.2 异常值检测算法阈值的确定

在传统的异常点检测方法中,往往存在选择阈值的困难,阈值的选择决定着检测异常点的准确性。该文将样本数据的均方根误差作为阈值,用预测的误差平方值与阈值进行比较,如果预测的误差平方值大于阈值,即为异常点。详细步骤如下:

step1:用建立好的 BP 神经网络模型对传入的数据 $x(n)(n = 1,2,\cdots,n)$ 进行预测,得到预测值 $\hat{x}(n)$。

step2:计算绝对误差 $\varepsilon(n) = |x(n) - \hat{x}(n)|$。

step3:计算 BP 神经网络数据样本的均方根误差作为阈值,计算公式如下所示:

$$\text{RMSD}(i) = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon(i)^2}{n}} \tag{13}$$

step4:计算 $P = \varepsilon(i)^2 - \text{RMSD}(i)$,如果 P 大于 0 则为异常点数据。

step5:用 BP 神经网络预测值代替异常点数据。

该方法解决了阈值选择的问题,提高了检测异常点的准确性,还为异常数据点找到了较好的替换数据。经过多次实验,该方法检测异常点准确率高,方便灵活。

3 实验测试

3.1 实验环境

软件环境:在 Windows10 操作系统下开发。

硬件环境:Inter Corei5 - 7400;4G 内存;500G 硬盘。

开发工具:Matlab R2014a;Visual Studio 2013。

3.2 实验数据集准备

实验选取录井真实数据作为实验训练样本,收集了 8 组实验数据。其中包括 4 组正常数据和 4 组异常数据,每组数据含同一井段钻压数据、钻速数据、扭矩数据、钻时数据,其异常数据基本信息如表 1 所示。将 8 组训练数据输入到建立好的 BP 神经网络模型进行训练,把训练好得到的权值系数、误差值等系数保存,并用保存的系数对测试数据进行计算判断。

表 1 异常数据集基本信息

异常数据集	数据量	异常值占比/%
第一组	500	8.4
第二组	1 000	7.5
第三组	1 500	10.6
第四组	2 000	9.8

选取大庆市某油田单井的 2 540 m ~ 2 590 m 录井数据进行测试,共选择钻压、钻速、扭矩、钻时数据各 500 条。其测试集数据基本信息如表 2 所示。

表 2 测试集数据

井深/m	钻压 (wob)	钻速 /rpm	扭矩 (turque)	钻时 (rop)
2 540.0	97.3	33	5.8	13.72
2 540.1	80.9	33	7.45	16.84
2 540.2	74.6	33	5	8.23
2 540.3	96	33	5.15	9.28
⋮				
2 589.6	55.6	33	5.25	17.66
2 589.7	60.7	33	4.75	13.99
2 589.8	56.8	33	5.65	7.03
2 589.9	53.4	33	5.2	15.91

3.3 评价指标

该文采用查全率(Recall)和误报率(FAR)作为评价指标来判断处理录井异常数据的好坏。查全率可反映查找正确异常点个数占总异常点个数的比例,其值越大,证明效果越好。误报率指正常数据被当作异常数据的比例,其值越小,效果越好。计算公式分别为:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{15}$$

式中,TP 为被预测为正的正样本;FN 为被预测为负的正样本;FP 为被预测为正的负样本;TN 为被模型预测为负的负样本。

3.4 结果分析

实验选取井深作为 X 轴坐标绘制图像,将测试集分别对格鲁布斯检验法、K-means 聚类算法、BP 神经网络进行异常数据检测,并绘制检测异常点情况图像,来对比 BP 神经网络检测异常点数据的效果。

图 2 和图 3 分别为格鲁布斯检验法和 K-means 聚类算法检测异常点的效果图。从图中可以看出,格鲁布斯检验法和 K-means 聚类算法均有异常点漏检或误检等情况。图 4 为 BP 神经网络模型的预测数据和实际检测样本的对比,可以看出异常点数据基本找出,并且异常点数据的预测值与实际数据更加紧密、

精准。

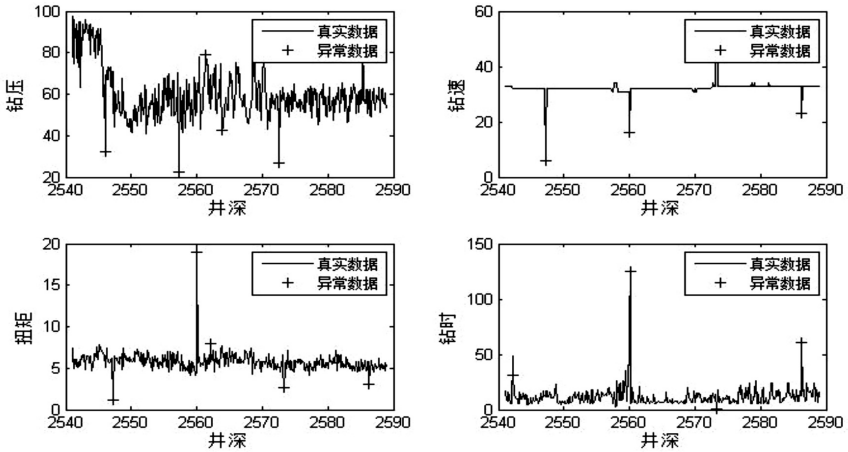

图 2　格鲁布斯检验法检测的异常点数据

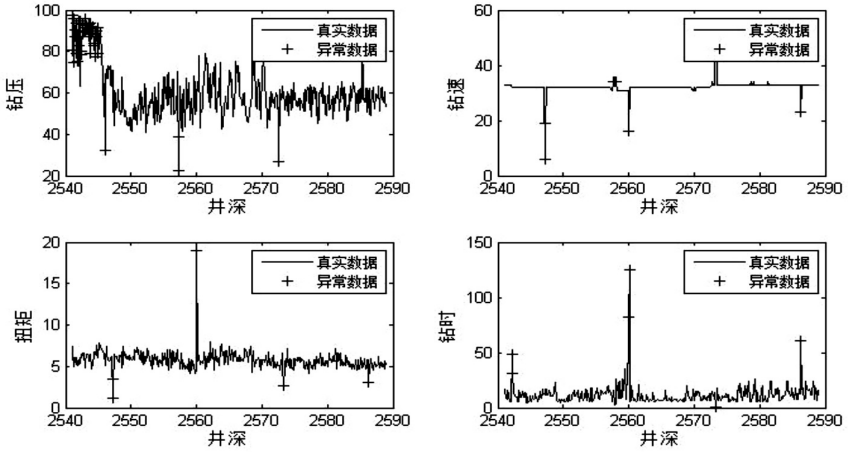

图 3　K-means 聚类算法检测的异常点数据

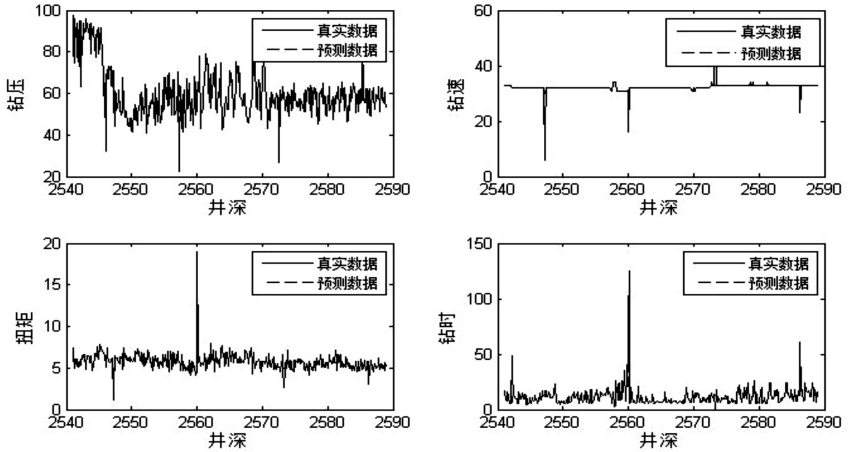

图 4　BP 神经网络检测样本值与预测值对比

由表 3 可知,针对录井异常数据情况,比较已有的异常检测算法,BP 神经网络方法检测异常点更加准确、误差更小。证明 BP 神经网络算法优于同类算法,可有效地检测大规模数据、异常聚集数据集中的异常样本,表现出了该算法的准确性和优越性。

表 3　查全率和误报率对比结果

数据集	算法	查全率/%	误报率/%
单井数据集	格鲁布斯检验法	83.2	3.6
	K-means 聚类算法	91.6	15.3
	BP 神经网络	98.4	0.8

4 结束语

针对“如何更快速、精准地处理录井数据中异常数据”的问题,提出了一种基于 BP 神经网络的异常值处理方法。通过建立 BP 神经网络模型,并计算 BP 神经网络数据样本的均方根误差作为阈值判断数据的异常情况。

由于录井数据的随机性,传统方法检测录井异常数据准确度不高,也无法对异常值数据实现替补,而基于 BP 神经网络的异常点数据检测模型能够有效检测异常点数据。经过异常数据处理后,不仅能够很好地把异常点数据处理掉,还能为异常点数据填补正常的数值,说明应用 BP 神经网络模型对异常点数据的处理具有较好的效果,结果比较可靠。

参考文献:

[1] 卓 琳,赵厚宇,詹思延. 异常检测方法及其应用综述[J]. 计算机应用研究,2020,37(S1):9-15.

[2] 朱嘉欣,包雨恬,黎 朝. 数据离群值的检验及处理方法讨论[J]. 大学化学,2018,33(8):58-65.

[3] 王 千,王 成,冯振元,等. K-means 聚类算法研究综述[J]. 电子设计工程,2012,20(7):21-24.

[4] 杨佳润. 数据挖掘之聚类分析算法综述[J]. 通讯世界,2017(16):291.

[5] 陈向东. 数据挖掘常用聚类算法分析与研究[J]. 数字技术与应用,2017(4):151-152.

[6] 陈钢花,胡 琮,曾亚丽,等. 基于 BP 神经网络的碳酸盐岩储层缝洞充填物测井识别方法[J]. 石油物探,2015,54(1):99-104.

[7] BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning,2009,2(1):1-127.

[8] 李 静,徐路路. 基于机器学习算法的研究热点趋势预测模型对比与分析——BP 神经网络、支持向量机与 LSTM 模型[J]. 现代情报,2019,39(4):23-33.

[9] MANICKAM R. Back propagation neural network for prediction of some shell moulding parameters[J]. Periodica Polytechnica Mechanical Engineering,2016,60(4):203-208.

[10] 刘金娣,李莉莉,高 静,等. 异常值检验方法的比较分析[J]. 青岛大学学报:自然科学版,2017,30(2):106-109.

[11] 曹志民,路成辉,刘 爽,等. 基于二分空间拉依达法的野值点剔除研究[J]. 化工自动化及仪表,2018,45(2):137-140.

[12] 林丽芬,肖 化,吴先球. 肖维勒准则和格拉布斯准则的比较[J]. 大学物理实验,2012,25(6):86-88.

[13] CHANDOLA V,BANERJEE A,KUMAR V. Anomaly detection:a survey[J]. ACM Computing Surveys,2009,41(3):1-58.

[14] SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al. Dropout:a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research,2014,15(1):1929-1958.

[15] 沈花玉,王兆霞,高成耀,等. BP 神经网络隐含层单元数的确定[J]. 天津理工大学学报,2008,24(5):13-15.

[16] SILVA A A,NETO I A L,MISSAGIA R M,et al. Artificial neural networks to support petrographic classification of carbonate-siliciclastic rocks using welllogs and textural information[J]. Journal of Applied Geophysics,2015,117:118-125.

(上接第 172 页)

[19] TSINGANOS P,SKODRAS A. On the comparison of wearable sensor data fusion to a single sensor machine learning technique in fall detection[J]. Sensors,2018,18(2):1-17.

[20] 吕 艳,张 萌,姜昊昊,等. 采用卷积神经网络的老年人跌倒检测系统设计[J]. 浙江大学学报:工学版,2019,53(6):1130-1138.

[21] 谷志瑜,刘建明,李建铎. 基于自回归模型和神经网络的跌倒检测算法[J]. 计算机工程与设计,2018,39(2):537-541.

[22] CATES B,SIM T,HEO H,et al. A novel detection model and its optimal features to classify falls from low-and high-acceleration activities of daily life using an insole sensor system[J]. Sensors,2018,18(4):1-16.

[23] 陈洪波,高 青,冯 涛,等. 基于足底压力信息的跌倒姿态聚类识别方法[J]. 电子技术应用,2016,42(5):113-115.

[24] 朱连杰,陈正宇,田晨林. 基于可穿戴设备的跌倒检测方法综述[J]. 计算机工程与应用,2019,55(18):8-14.

[25] 忽丽莎,王素贞,陈益强. 基于可穿戴设备的跌倒检测算法综述[J]. 浙江大学学报:工学版,2018,52(9):1717-1728.

[26] 胡双杰,秦建邦,郭 薇. 基于特征自动提取的跌倒检测算法[J]. 传感技术学报,2018,31(12):1842-1847.

[27] 焦李成,杨淑媛,刘 芳,等. 神经网络七十年:回顾与展望[J]. 计算机学报,2016,39(8):1697-1717.

[28] 刘 勃,王明伟,常立博. 老年人运动状态监测和跌倒报警系统[J]. 计算机与现代化,2020(5):101-105.

[29] 王晓雷,李栋豪,郑晓婉,等. 基于 RBF 神经网络的跌倒检测算法研究[J]. 电子测量与仪器学报,2019,33(11):185-191.

[30] AZIZ O,MUSNGI M,PARK E J,et al. A comparison of accuracy of fall detection algorithms (threshold-based vs. machine learning) using waist-mounted tri-axial accelerometer signals from a comprehensive set of falls and non-fall trials[J]. Medical & Biological Engineering & Computing,2017,55(1):45-55.