

DOI:10.3969/j.issn.1005-202X.2023.08.019

医学生物信息

## 基于无创血压构建 MIMIC-III 波形数据库匹配子集

陈智恒<sup>1</sup>, 尹存芳<sup>2</sup>, 罗俊卿<sup>3</sup>

1. 南方医科大学公共卫生学院, 广东 广州 510515; 2. 南方医科大学南方医院急诊科, 广东 广州 510515; 3. 南方医科大学生物医学工程学院, 广东 广州 510515

**【摘要】**为解决 MIMIC-III 临床数据库与波形数据库匹配度低的问题,提出一种基于相似度匹配的算法以构建全新的 MIMIC-III 波形数据库匹配子集。首先利用 BP 神经网络,对波形数据库缺失的 ICU 位置信息进行填补。然后利用匹配算法计算两个数据库无创血压的相似度。最后检验并选取相似度高的匹配数据,得出新的波形数据库匹配子集。本算法共匹配了 1 133 条波形数据库数据。匹配结果表明,该方法匹配的平均正确率为 89.03%,具有较高的准确度和效率。这一方法为研究人员构建 MIMIC 数据库提供了新的思路,并为基于 MIMIC 数据库的临床研究提供更多的信息。

**【关键词】**重症医学;无创血压;临床数据库;波形数据库

**【中图分类号】**R318;R459.7

**【文献标志码】**A

**【文章编号】**1005-202X(2023)08-1039-06

## Construction of a MIMIC-III waveform database matched subset based on non-invasive blood pressure

CHEN Zhiheng<sup>1</sup>, YIN Cunfang<sup>2</sup>, LUO Junqing<sup>3</sup>

1. School of Public Health, Southern Medical University, Guangzhou 510515, China; 2. Department of Emergency, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China; 3. School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China

**Abstract:** A similarity matching algorithm is proposed to construct a MIMIC-III waveform database matched subset for solving the problem of low matching degree between MIMIC-III clinical database and waveform database. After filling in the missing ICU location information using BP neural network, the matching algorithm is used to calculate the similarity of non-invasive blood pressure between the two databases, and the matching records with high similarity are examined and selected to construct a new matched subset. A total of 1 133 waveform database records are matched. The matching results show that the proposed method has high accuracy and efficiency, with an average matching accuracy of 89.03%. The method provides new ideas for researchers to construct MIMIC database and provides more information for clinical research based on MIMIC database.

**Keywords:** critical care medicine; non-invasive blood pressure; clinical database; waveform database

### 前言

重症监护室中监护仪数据可以帮助医生及时了解患者信息,并及时改进治疗方案<sup>[1]</sup>。虽然监护室的监护仪会产生数量级极为庞大的信息,但在大多数情况下,只有一小部分信息被充分利用<sup>[2]</sup>。尽可能利

用监护仪产生的数据将进一步推动重症医学的发展。

MIMIC-III (Medical Information Mart for Intensive Care)是一个大型单中心数据库,其包含了马萨诸塞州波士顿贝斯以色列女执事医疗中心在 2001 年到 2012 年期间收集的重症监护室数据。这些数据已广泛应用于临床各领域的研究<sup>[3-5]</sup>。此外, MIMIC-III 数据库拥有一个与之对应的 MIMIC-III 波形数据库,该数据库包含与 MIMIC-III 数据库相同患者群体的波形数据<sup>[3, 5-6]</sup>。为符合医学伦理要求,两个数据库分别进行了去识别化处理。因此,完成上述两个数据库的匹配工作将会变得十分困难。

**【收稿日期】**2023-05-26

**【作者简介】**陈智恒,研究方向:重症医学, E-mail: czh0929\_rock@163.com

**【通信作者】**罗俊卿,博士,实验师,研究方向:重症医学、医学信息化, E-mail: luojunqing@gmail.com

尽管 Moody 等已经对 MIMIC-III 临床数据库与 MIMIC-III 波形数据库完成了 22 317 条数据的匹配工作<sup>[3,5,7]</sup>,但仍存在配套数据匹配完成度低的问题。并且近 10 年来有关 MIMIC 数据库的研究不断增多<sup>[8]</sup>,相关研究可能出现不完整或不准确的结果。因此急需对临床数据库与波形数据库剩余数据进行深度匹配。本文旨在提出一种相似度匹配算法,对 MIMIC-III 数据库及其波形数据库进行匹配并得出一个全新的波形数据库子集。

## 1 方法简介

### 1.1 相似度匹配算法

相似度分析是许多计算机科学和统计学领域中的重要问题,它涉及比较和度量不同对象之间的相似程度。欧几里德距离(欧氏距离)作为一种经典的距离度量方法,具有计算简单、直观易懂,适用于各种数据类型和特征向量表示的优点,能够有效地衡量对象之间的差异性<sup>[9]</sup>。

欧氏距离最初用于计算欧几里德空间两个点的距离。假设  $x,y$  为  $n$  维空间的两个点,它们之间的距离为<sup>[10-11]</sup>:

$$d(x,y)=\sqrt{(\sum(x_i-y_i)^2)}$$

(1)

当用欧氏距离表示相似度,用以下公式转换:

$$\text{sim}(x,y)=\frac{1}{1+d(x,y)}$$

(2)

由上述公式可见,欧氏距离法关注两点之间的总体距离。由于欧式距离之间的标量总和和相关,此方法能较好地把握总体距离。

### 1.2 BP神经网络算法

BP 神经网络(Backpropagation Neural Network)是一种由非线性变换单元构成的前馈性神经网络,具有优良的非线性表征能力<sup>[12-13]</sup>,广泛运用于预测与分类工作。常见的 BP 神经网络模型如图 1 所示,共有 3 层,包含输入层、隐藏层与输出层。层与层之间的神经元全连接,神经元的输入来自上一层的输出<sup>[14-15]</sup>。此类型的神经网络可逼近几乎所有的连续函数<sup>[16]</sup>。数据的分类与预测主要分为正向传播与反向误差传播过程<sup>[17]</sup>。后者可以根据输入层信息与输出层信息之间的误差不断调整权重,令 BP 神经网络能反映复杂的变量关系<sup>[18]</sup>。

## 2 构建 MIMIC-III 波形数据库匹配子集

### 2.1 基于 BP 神经网络填补 ICU 信息

ICU 位置信息是联系 MIMIC-III 波形数据库与 MIMIC-III 临床数据库最为重要的依据之一。然而,

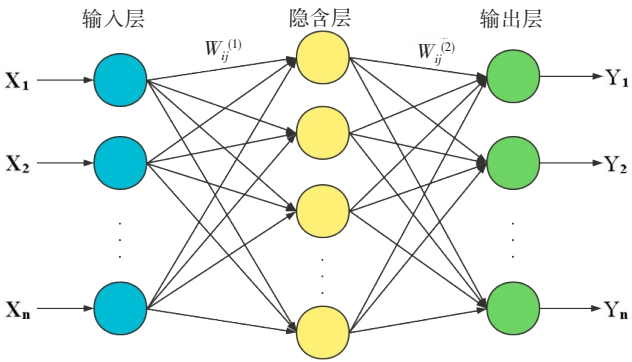


图1 BP神经网络结构示意图  
Figure 1 BP neural network structure

在 67 380 条识别化数据中,有 6 558 条数据缺失 ICU 位置信息,这可能是在 MIMIC-III 波形数据库去识别化过程中或者监护仪收集信息时发生配置错误导致的。不同 ICU 科室的监护仪会记录患者不同的生理信号,因此可以利用已知 ICU 信息的数据对缺失 ICU 信息的数据进行填补。运用 BP 神经网络进行数据填补的流程图如图 2 所示。

选取 MIMIC-III 波形数据库及其匹配子集中已知 ICU 位置信息的 52 289 条数据作为样本。经过图 2 所示的预处理后,用前 95% 的数据集作为训练集,后 5% 的数据集为测试集进行 BP 神经网络训练。BP 神经网络的参数如下:波形数据的 183 种生理信号作为输入层,ICU 类别为输出层。由于隐含层的复杂度过高会导致过拟合现象<sup>[19-20]</sup>,因此隐含层的神经元数量应参照经验公式设定为 18 个<sup>[21]</sup>。网络初始学习速率为 0.1,训练次数为 1 000 次,目标最小误差为 0.000 01。利用均方根误差(RMSE)评估训练效果<sup>[22-23]</sup>。当训练误差大于目标最小误差时,网络通过反向传播去不断调整各权重,直至误差小于目标最小误差或达到训练次数后完成训练。把测试集代入现有训练完成的 ICU 分类模型中进行验证。可得各 ICU 位置信息的累积误差曲线如图 3 所示。

由图 3 可知,除外科重症监护室(SICU)的累积误差曲线较为平缓外。其余的 ICU 误差大部分集中在 0.1 以内,说明 BP 神经网络训练的分类模型具有一定的分类能力。然而,不同 ICU 类别的数据可能出现近乎相同的生理信号种类,这将会导致部分数据的 ICU 分类模糊。需要利用特异性生理信号对分类结果进行检验与重新分类,如血气分配系数(BAP)、毛细血管二氧化碳分压(CPCO<sub>2</sub>)等。具体分类结果见网站文件 classification(<https://github.com/9rockchen/MIMIC-III-waveform-new-subset.git>)。

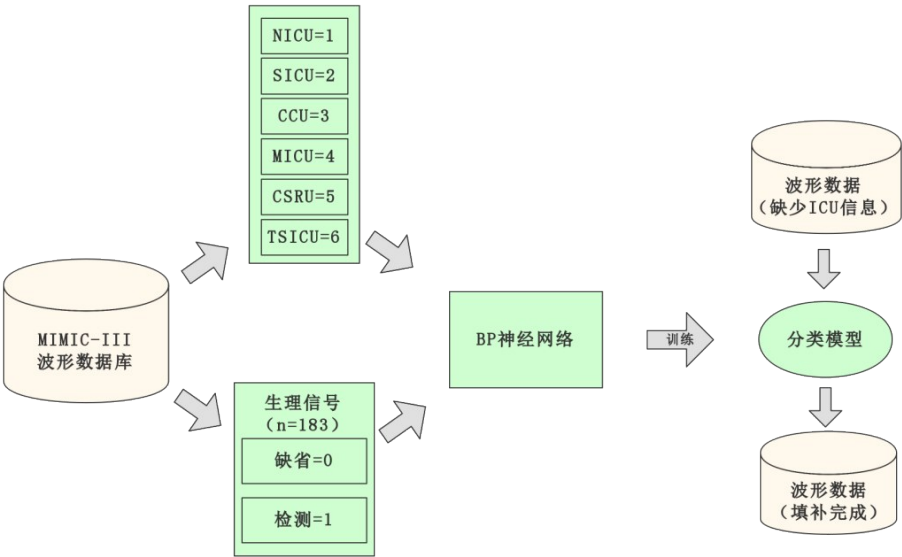


图2 填补流程图

Figure 2 Flowchart of data filling

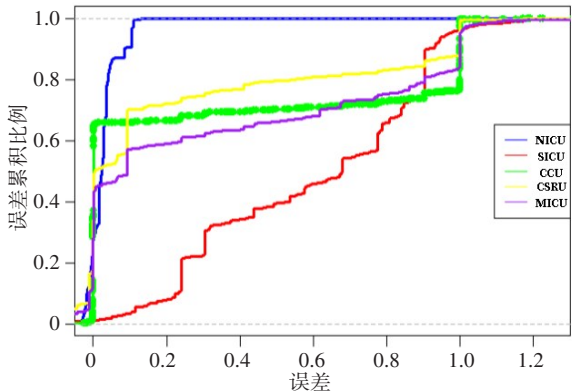


图3 ICU 累积误差分布图

Figure 3 Cumulative error distribution of ICU

NICU:新生儿重症监护室,SICU:外科重症监护室,CCU:冠状动脉护理单元,CSRU:心脏手术恢复监护室,MICU:医学重症监护室

2.2 基于相似度算法匹配波形数据

结合 1.1 节信息填补的结果,对未匹配的波形数据进行相似度匹配的流程如图 4 所示。总结起来此方法的基本思路涵盖以下 4 个步骤。

(1) 数据校验。MIMIC-III 临床数据库的 transfer 表主要记录了每个患者出入 ICU 的时间与 ICU 科室信息。在 SUBJECT\_ID 相同的情况下,将 MIMIC-III 匹配子集中每条数据的时间信息与 transfer 表中出入 ICU 科室的时间进行对比,剔除 MIMIC-III 波形数据库匹配子集中无法与临床数据库出入 ICU 科室时间相对应的数据。把剔除出来的数据与未匹配的数据整合形成新的待匹配的波形数据集。

(2)提取数据库无创血压信息。利用 Python3.10 中 wfdb 库的 p\_signal 函数提取出一条未匹配的波形

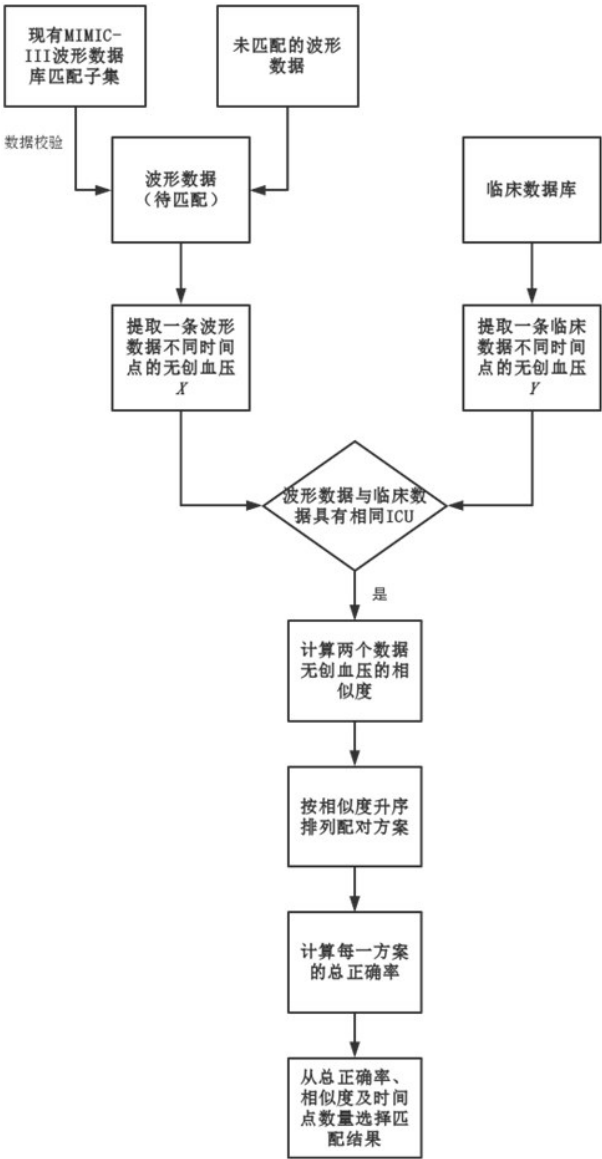


图4 数据匹配流程图

Figure 4 Flowchart of data matching

数据库数据的无创血压值及对应的时间信息Y。并且在ICU类型相同的情况下,利用PostgreSQL 14提取临床数据库chartevent表中一条临床数据的无创血压及其对应的时间信息X。

(3)计算两个数据无创血压的相似度。运用欧氏距离法计算两个数据无创血压的欧氏距离,并换算成相似度。每次计算相似度将形成一个匹配方案。按照相似度降序的方式对配对方案进行排列,形成配对方案集。

(4)匹配校验。提取第(3)步中的配对方案集,并对前25%的方案集进行校验。对比相同时间点下两个数据无创血压的数值。如果数值相同,则此时时间点的无创血压匹配正确。把所有匹配正确的时间

点个数除以总的时间点个数,得出总正确率。选取无创收缩压与平均压的正确率均大于75%、相似度较高且涵盖时间点数量多于3个的配对方案作为匹配结果。

### 3 结果

#### 3.1 数据校验结果

经过数据清洗,总共剔除3 913条与MIMIC-III临床数据库出入ICU时间无法对应的波形数据库数据。具体剔除结果详见网站(<https://github.com/9rockchen/MIMIC-III-waveform-new-subset.git>)。上述数据清洗的过程以SUBJECT\_ID为125的实例进行说明,如表1所示。

表1 MIMIC-III波形数据库匹配子集校验表  
Table 1 MIMIC-III waveform database matched subset checklist

MIMIC-III 临床数据库			MIMIC-III 波形数据库子集		
SUBJECT_ID	INTIME	OUTTIME	SUBJECT_ID	INTIME	OUTTIME
125	2179/2/14 21:23:00	2179/2/14 21:28:00	125	2187/01/27 14:22:00	2187/01/28 05:28:00
125	2179/2/14 21:28:00	2179/2/16 7:34:00	125	2187/01/28 07:29:00	2187/01/31 20:20:00
125	2179/2/16 7:34:00	2179/2/16 9:36:00	125	2187/02/13 04:17:00	2187/02/13 16:22:00
125	2179/2/16 9:36:00	2179/2/17 15:05:00	125	2187/02/13 18:46:00	2187/02/14 10:02:00
125	2179/2/17 15:05:00	2179/2/21 18:04:00	125	2187/02/14 11:28:00	2187/02/14 16:15:00
125	2179/2/21 18:04:00	Null	125	2187/02/14 20:18:00	2187/02/20 21:20:00

左侧3列数据为MIMIC-III临床数据库中SUBJECT\_ID为125的患者所有出入ICU的记录。右侧数据为MIMIC-III波形数据库匹配子集同一患者的记录。两个数据库的脱敏化日期与时间完全无法对应。把诸如此类的数据从MIMIC-III波形数据库匹配子集剔除并加入未匹配波形数据库子集中重新进行匹配。

#### 3.2 匹配结果

运用上述所提出的算法,最终共有1 133条MIMIC-III波形数据成功与MIMIC-III临床数据库匹配。共覆盖患者859名。其平均正确率为89.03%。以SUBJECT-ID为57678的患者与波形编号为3366760的波形数据作为示例进行说明。

在ICU类别相同的前提下,计算所有未匹配波形数据与临床数据之间的欧氏距离。换算为相似度并降序排列,选取前25%的配对方案进行校验。校验的评判指标为总正确率、相似度及涵盖时间点数量。筛选出最为合理的波形记录,其波形编号为3366760。此波形数据无创收缩压与平均压的欧氏

距离相似度分别为0.161与0.196,并且覆盖了37个时间点。详细相似度结果见网站附件odnbp。验证结果如图5所示。

由图5可得,由于两个数据库采用去识别化处理,x轴的数值表示从第0s开始到记录结束的时间点,y轴为某一时间点无创收缩压/平均压的具体值。蓝点为MIMIC-III波形数据库以1s为周期记录的无创收缩压/平均血压的值。红点为在同一时刻临床数据库记录无创血压的值。红点与蓝线重合时,说明当前时间点临床数据库记录的无创平均血压/收缩压数值与波形数据库记录的无创平均血压/收缩压相等,此时间点匹配正确。无创平均血压与收缩压分别共有37个值作为验证点。其中,平均血压共有34个点验证匹配正确,正确率为91.89%;收缩压共有35个点验证匹配正确,正确率为94.59%。两个生理指标的匹配正确率都在90%以上,在误差允许范围内,可以认为编号为3366760的波形数据与SUBJECT-ID为57678的患者匹配成功。



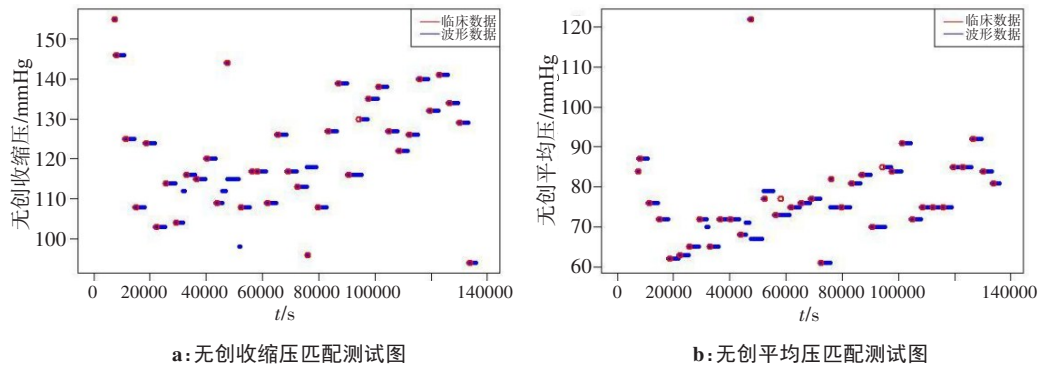


图5 编号3366760波形数据匹配测试

Figure 5 No.3366760 waveform data matching test

## 4 讨论

本研究旨在开发相似度算法,对现有的MIMIC-III波形数据库匹配子集进行校验并利用无创血压构建新的匹配子集。最终成功完成了1133条波形记录的匹配,涵盖了859名患者。然而,值得注意的是,本研究中所提出的算法仅利用无创血压对尚未匹配的波形记录进行了深度匹配。事实上,仍有其他具有足够特异性的指标,例如有创血压,可用于完成未匹配的波形记录匹配工作。未来的研究重点为充分挖掘这些特异性指标,以进一步完善对MIMIC-III波形数据库与临床数据库的匹配工作。

## 5 数据库匹配子集使用说明

整理符合结果验证的1133条波形记录,构建出新的MIMIC-III波形数据库子集(<https://github.com/9rockchen/MIMIC-III-waveform-new-subset.git>)。仍以SUBJECT-ID为57678的患者为示例作为子集的使用说明:(1)SUBJECT-ID为57678的患者的所有信息都放在文件名为57678的文件夹中。(2)为降低数据库大小,本数据库匹配子集仅包含索引文件,汇总的头文件与对应的波形记录文件。57678共包含3个文件,分别为m057678-2182-06-14-21-58-31n.dat、m057678-2182-06-14-21-58-31n.he、3366760record。其中,dat文件为波形数据文件,n.he文件为数字记录的头文件,文件名带有record的文件为索引文件,若需要分段的波形文件或头文件,可通过此文件索引MIMIC-III波形数据库的相应子文件。(3)新的匹配子集m057678-2182-06-14-21-58-31记录了从2182年6月14日21:58:31开始,长度为138928s的波形信息。与MIMIC-III数据库一致,新的匹配子集的日期随机移动到未来若干天实现去识别化,以保证患者权益。(4)有关原匹配子集的具体内容与格式请参照WFDB的应用指南与其原文献<sup>[24-25]</sup>。

## 【参考文献】

- [1] Kelly FE, Fong K, Hirsch N, et al. Intensive care medicine is 60 years old: the history and future of the intensive care unit[J]. Clin Med (Lond), 2014, 14(4): 376-379.
- [2] Celi LA, Mark RG, Stone DJ, et al. "Big data" in the intensive care unit. Closing the data loop[J]. Am J Respir Crit Care Med, 2013, 187(11): 1157-1160.
- [3] Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database[J]. Sci Data, 2016, 3: 160035.
- [4] Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database (version 1.4). PhysioNet[EB/OL]. <https://doi.org/10.13026/C2XW26>. 2016.
- [5] Goldberger AL, Amaral LA, Glass LA, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals[J]. Circulation, 2000, 101(23): E215-220.
- [6] Moody B, Moody G, Villarroel M, et al. MIMIC-III Waveform Database (version 1.0). PhysioNet[EB/OL]. <https://doi.org/10.13026/c2607m>. 2020.
- [7] Moody B, Moody G, Villarroel M, et al. MIMIC-III Waveform Database Matched Subset (version 1.0). PhysioNet[EB/OL]. <https://doi.org/10.13026/c2294b>. 2020.
- [8] 张心怡, 张六一. MIMIC数据库研究现状与热点分析[J]. 医学信息, 2022, 35(1): 19-23.  
Zhang XY, Zhang LY. Research status and hot spot analysis of MIMIC database[J]. Medical Information, 2022, 35(1): 19-23.
- [9] 黄晓俊, 李晓宏, 郭战魁, 等. 基于相关系数的一致性配组法[J]. 电子工艺技术, 2023, 44(2): 12-16.  
Huang XJ, Li XH, Guo ZK, et al. Consistency grouping method based on correlation coefficient[J]. Electronics Process Technology, 2023, 44(2): 12-16.
- [10] 董旭, 魏振军. 一种加权欧氏距离聚类方法[J]. 信息工程大学学报, 2005, 6(1): 23-25.  
Dong X, Wei ZJ. A clustering method of euclid distance with weights[J]. Journal of Information Engineering University, 2005, 6(1): 23-25.
- [11] 朱仁治. 一种加权欧氏距离聚类算法的改进[J]. 计算机与数字工程, 2016, 44(3): 421-424.  
Zhu LZ. Improvement of weighted euclidean distance clustering algorithm[J]. Computer & Digital Engineering, 2016, 44(3): 421-424.
- [12] Panwar M, Gautam A, Biswas D, et al. PP-Net: a deep learning framework for PPG-based blood pressure and heart rate estimation[J]. IEEE Sens J, 2020, 20(17): 10000-10011.
- [13] 杨延璞, 余隋怀, 陈登凯. 运用遗传算法的产品造型设计方案优化方法[J]. 现代制造工程, 2012(3): 127-131.  
Yang YP, Yu SH, Chen DK. Product design optimization method based on genetic algorithm[J]. Modern Manufacturing Engineering, 2012(3): 127-131.
- [14] 刘梦尧. 基于BP神经网络的量化选股模型应用研究[D]. 长春: 长春工业大学, 2022.  
Liu MY. Research on the application of multi-factor quantitative stock selection model based on BP Neural network [D]. Changchun: Changchun University of Technology, 2022.
- [15] 周中, 邓卓湘, 陈云, 等. 基于GA-BP神经网络的泡沫轻质土强度

- 预测[J]. 华南理工大学学报(自然科学版), 2022, 50(11): 125-132.  
Zhou Z, Deng ZX, Chen Y, et al. Strength prediction of foam light soil based on GA-BP neural network [J]. Journal of South China University of Technology (Natural Science Edition), 2022, 50(11): 125-132.
- [16] 李达, 陈达, 江新标, 等. 基于BP神经网络从活化数据中求解1 MeV等效中子注量[J]. 现代应用物理, 2015, 6(4): 248-253.  
Li D, Chen D, Jiang XB, et al. Calculation of 1 MeV equivalent neutron fluence from activation data based on BP neural networks[J]. Modern Applied Physics, 2015, 6(4): 248-253.
- [17] Huang B, Chen W, Lin CL, et al. MLP-BP: a novel framework for cuffless blood pressure measurement with PPG and ECG signals based on MLP-Mixer neural networks[J]. Biomed Signal Process Control; 2022, 73: 103404.
- [18] Kurylyak Y, Lamona F, Grimaldi D. A neural network based method for continuous blood pressure estimation from a PPG signal[C]//2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 2013.
- [19] 许扬, 蔡安民, 吴梓秋, 等. 基于BP神经网络和多因素权重分析的气热除冰温度影响因素研究[J]. 热力发电, 2022, 51(12): 131-140.  
Xu Y, Cai AM, Wu ZQ, et al. Influencing factors of air thermal deicing temperature based on BP neural network and multi-factor weight analysis[J]. Thermal Power Generation, 2022, 51(12): 131-140.
- [20] 许凯文. 基于遗传算法优化BP神经网络的深基坑地连墙变形预测[J]. 粉煤灰综合利用, 2021, 35(5): 6-11.  
Xu KW. Deformation prediction of diaphragm wall of deep foundation pit based on BP neural network improved by genetic algorithm[J]. Fly Ash Comprehensive Utilization, 2021, 35(5): 6-11.
- [21] 程志义, 周广浩, 程炜晴, 等. 基于BP神经网络的激光叠焊焊接接头熔深预测研究[J]. 城市轨道交通研究, 2020, 23(2): 141-144.  
Cheng ZY, Zhou GH, Cheng WQ, et al. Research on laser lap weld depth prediction based on BP neural network[J]. Urban Mass Transit, 2020, 23(2): 141-144.
- [22] 张伟娜. 基于深度学习与矩阵分解的推荐算法研究[D]. 广州: 华南理工大学, 2020.  
Zhang WN. Research on recommendation algorithm based on deep learning and matrix factorization [D]. Guangzhou: South China University of Technology, 2020.
- [23] 汪子豪, 秦其明, 孙元亨, 等. 基于BP神经网络的地表温度空间降尺度方法[J]. 遥感技术与应用, 2018, 33(5): 793-802.  
Wang ZH, Qin QM, Sun YH, et al. Downscaling of remotely sensed land surface temperature with the BP neural network [J]. Remote Sensing Technology and Application, 2018, 33(5): 793-802.
- [24] Moody GB. WFDB Applications Guide[Z]. 2019.
- [25] Moody GB. The WFDB Software Package[Z]. 2018.
- (编辑: 薛泽玲)