

引用格式:张网娟,许国艳,李敏佳,等. 基于卷积神经网络的缺失数据填充方法 [J]. 微电子学与计算机,2019,36(3):48-52.
[ZHANG Wang—juan, XU Guo—yan, LI Min—jia, et al. Missing data imputation approach based on convolutional neural network[J]. Microelectronics & Computer,2019,36(3):48-52.]

基于卷积神经网络的缺失数据填充方法

张网娟,许国艳,李敏佳,朱 帅

(河海大学 计算机与信息学院,江苏 南京 211100)

摘 要: 受不利因素影响,传感器网络中的数据不可避免地存在缺失,阻碍上层应用对数据的分析与处理. 使用能够自主提取和学习有效特征结构的卷积神经网络对缺失数据问题进行研究,提出一种基于卷积神经网络的缺失数据填充方法. 首先,分别针对时间序列数据的时间相关性和传感器节点间的空间相关性,使用卷积神经网络填充模型对缺失数据进行单维度相关性的填充. 然后,根据时间维度和空间维度的填充结果,考虑时空相关性对缺失数据进行填充. 最后,在真实数据集上进行实验验证,实验结果表明考虑时空多维度相关性的填充结果优于只考虑单维度相关性的填充结果,并与 BP 模型进行对比,验证了卷积神经网络填充模型的有效性.

关键词: 缺失数据;填充;卷积神经网络;时空相关性

中图分类号: TP311

文献标识码: A

文章编号: 1000—7180(2019)03—0048—05

Missing data imputation approach based on convolutional neural network

ZHANG Wang-juan, XU Guo-yan, LI Min-jia, ZHU Shuai

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: In the sensor network, the missing data is inevitable due to many of interference factors, which hinders data analysis and processing. In this research, by using the convolution neural network(CNN) which can extract and learn effective features data independently, a missing data imputation approach based on convolution neural network was proposed. Firstly, for the temporal correlation of time series and the spatial correlation between sensor nodes, the CNN imputation model was used to estimate the missing data with single dimensional correlation. Then, based on the imputation results of temporal dimension and spatial dimension, the missing data were imputed by considering the multidimensional correlation. Finally, experimental verification is performed on the real data set. Experiment results show that the imputation result considering the temporal and spatial multidimensional correlation is better than the imputation result considering the single dimensional correlation. And compared with the BP model, the validity of the convolution neural network imputation model is verified.

Key words: missing data; imputation; convolution neural network(CNN); temporal and spatial correlation

1 引言

随着计算机和通信技术、传感器技术及嵌入式

技术的迅速发展,传感器设备被大量运用于军事、医疗、工业等多个领域. 传感器节点采集的数据能帮助人们更好地了解监测区域,然而由于设备软硬件故

障、网络传输问题及工作环境恶劣等不可控因素,传感器网络中的数据存在不同程度的缺失,降低了数据的可用性^[1].同时,现有的数据分析工具大多是基于完整数据集提出的,无法直接处理含有缺失数据的不完整数据集,因此如何有效处理缺失数据成为当前需要解决的问题.

直接丢弃缺失数据简单易行,但当缺失数据比例较大时,此方法会造成原始数据失真,损害推理能力.因此,对缺失数据进行填充是一种更合理且非常必要的方法.对于缺失数据填充方法的研究,前人已有大量的研究成果.均值填充是典型的基于统计的方法,此方法可保持序列的均值不变,但填充精度较低. K 最近邻填充法通过 K 个最近邻居来估计缺失数据,但 K 值的选择限制了该方法的实用性.针对传感器网络数据的特点,文献[2]通过属性约简划分属性的重要性,再根据属性的重要程度选择相应的填充算法.文献[3]考虑数据的时间相关性,使用局部时间索引(Local Time Index, LTI)策略对时间序列中存在的缺失数据进行填充.文献[4]考虑数据时空相关性,使用概率主元分析(Probabilistic Principal Component Analysis, PPCA)和混合概率主元分析(Mixed Probabilistic Principal Component Analysis, MPPCA)论证了同时使用时间和空间信息可以提高填充的准确性.

此外,深度学习模型^[5-7]通过其非线性和自学习的特点可对非线性函数进行逼近,对缺失数据进行估计和填充,如使用BP神经网络估计缺失值^[6-7],但考虑到参数学习效率,实际应用中模型的构建一般不超过三层^[8].卷积神经网络(Convolutional Neural Network, CNN)是一种常用的多层深度学习模型,通过卷积和池化操作可以提取低层的局部特征,然后在更高层综合局部特征得到全局特征.自2012年Krizhevsky等人使用扩展深度的CNN在ImageNet图像分类竞赛中取得最佳分类效果,使得CNN越来越受到研究者的关注.当前,CNN在图像分类、人脸特征提取、图像语义分割等多个方面表现突出,已有研究表明,CNN在数值回归问题上也有较好的效果^[9].

基于以上理论基础,本文考虑将数据的时空相关性与深度学习模型相结合,提出一种基于卷积神经网络的缺失数据填充方法.

2 相关工作

2.1 卷积神经网络

卷积神经网络^[8]的基本结构由输入层、卷积层

(Convolutional Layer)、池化层(Pooling Layer,也称为下采样层)、全连接层和输出层组成.在卷积神经网络中,采用卷积层和池化层交替设置的方式进行模型的构造,通过对多维矩阵进行逐层特征提取,学习数据内部的信息.与BP神经网络相比,虽然它们拥有相似的训练方式,但全连接神经网络的深度较浅,而CNN的局部连接、权值共享和池化操作,可减少连接数目和训练参数,增加神经网络的训练深度且易于模型的训练.

卷积层包含多个特征面,每个特征面中的神经元通过卷积核与上一层的特征面进行局部相连接,使用非线性函数对局部加权并进行激活,从而获得卷积层中每个神经元的输出值.卷积层可描述为:

$$\mathbf{H}_{i+1} = f(\mathbf{A}_i \otimes \mathbf{W}_{i+1} + b_{i+1}) \quad (1)$$

式中, $\mathbf{H}_{i+1}$ 为卷积层的输出特征矩阵; $\mathbf{A}_i$ 为原始输入矩阵或上一层池化层的输出特征矩阵; $\mathbf{W}_{i+1}$ 为第 $i+1$ 层卷积核权重向量; $\otimes$ 表示进行卷积操作; b_{i+1} 为第 $i+1$ 层的偏置; $f(\cdot)$ 为非线性激活函数.

池化层同样包含多个特征面,其每一个特征面与卷积层的一个特征面唯一对应,因此池化操作不会改变特征面的个数.池化层是采用一定的下采样操作对特征图进行二次提取特征的过程,常用的池化法有最大池化、均值池化和随机池化.池化层可描述为:

$$\mathbf{A}_i = f_{\text{sub}}(\mathbf{H}_{i-1}) \quad (2)$$

式中, $\mathbf{A}_i$ 为池化层的输出特征矩阵; $\mathbf{H}_{i-1}$ 为第 $i-1$ 层卷积层的输出特征矩阵,作为池化层的输入特征矩阵; $f_{\text{sub}}(\cdot)$ 为下采样函数.

CNN以最小化损失函数 $L(W, b)$ 为网络的训练目标,本文选用均方误差(Mean Squared Error, MSE)作为损失函数:

$$\text{MSE}(W, b) = \frac{1}{N} \sum_{i=1}^N (Y(i) - \hat{Y}(i))^2 \quad (3)$$

式中, $Y(i)$ 表示期望值, $\hat{Y}(i)$ 表示卷积神经网络的输出值.

经多个卷积层和池化层的交替传递,逐层提取输入矩阵的特征后,通过全连接网络对提取的多个特征进行整合.

2.2 二维信息矩阵

卷积神经网络多用于图像分类领域, $m \times n$ 的图像数据矩阵能够很好地满足卷积层和池化层的要求,但本文的原始数据样本并不是以矩阵方式存储的,因此为了满足CNN卷积层和池化层的操作要求,将原始数据向量转换为二维信息矩阵^[9].

设 x 为原始数据向量,且为列向量,则二维信息矩阵为:

$$\mathbf{X} = xx^T = \begin{pmatrix} a_1^2 & a_1 a_2 & \cdots & a_1 a_n \\ a_2 a_1 & a_2^2 & \cdots & a_2 a_n \\ \cdots & \cdots & \cdots & \cdots \\ a_n a_1 & a_n a_2 & \cdots & a_n^2 \end{pmatrix} \quad (4)$$

使用原始数据向量构建的二维信息矩阵,不仅保持了原始数据向量的数据特征,同时满足了 CNN 对输入数据的要求。

3 基于卷积神经网络的缺失数据填充方法

3.1 问题的形式化定义

设传感器网络中有 N 个数据源 $S = \{s_1, s_2, \dots, s_N\}$,第 $n(n \in N)$ 个传感器节点的时间序列为 $s_n = \{d_{n1}, d_{n2}, \dots, d_{nT}\}$ (共 T 个时间步),第 $t(t \in T)$ 个时刻 N 个传感器节点的数据序列为 $D_t = \{d_{1t}, d_{2t}, \dots, d_{Nt}\}$. 本文研究的问题可定义为:在传感器网络中,传感器节点 n 在 t 时刻数据发生缺失,使用传感器节点 n 的时间序列数据和监测区域内其他传感器节点的数据对 t 时刻的缺失值进行估计和填充。

3.2 缺失数据填充方法

使用卷积神经网络对缺失数据进行填充有三种基本思路:时间相关性、空间相关性和时空相关性。

3.2.1 时间相关性

基于时间相关性的填充只利用缺失数据所在传感器节点的时间序列数据进行 CNN 回归,填充缺失数据. 本文采用文献[3]的局部时间索引(LTI)策略构建原始训练集和测试集,具体方法如下:

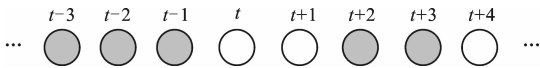


图1 含有缺失数据的时间序列

图1是一个含有缺失数据的时间序列,其中 d_t, d_{t+1}, d_{t+4} 为缺失数据,记录上方的数值代表全局时间索引,以窗口大小为 2 来建立基于时间相关性填充的原始训练集和测试集. 非缺失数据被用于构建训练集,缺失数据用于构建测试集。

对于非缺失数据 d_{t-1} ,它的前两个非缺失数据为 d_{t-3}, d_{t-2} ,对应的全局时间索引为 $t-3$ 和 $t-2$,由于第一个非缺失数据下标为 $t-3$,因此局部时间索引表示为 $\{0, 1, 2\}$,因此 d_{t-1} 所构建的原始训练向量为 $p_{t-1} = (\{d_{t-3}, d_{t-2}, 0, 1, 2\}, \{d_{t-1}\})$. 以此类推, d_{t+2}, d_{t+3} 构建的原始训练向量分别为 $p_{t+2} = (\{d_{t-2}, d_{t-1}, 0, 1, 4\}, \{d_{t+2}\})$ 和 $p_{t+3} = (\{d_{t-1}, d_{t+2}, 0, 3,$

$4\}, \{d_{t+3}\})$.

对于缺失数据 d_t ,它的前两个非缺失数据为 d_{t-2}, d_{t-1} ,对应的全局时间索引为 $t-2$ 和 $t-1$,由于第一个非缺失数据下标为 $t-2$,因此局部时间索引为 $\{0, 1, 2\}$,因此 d_t 所构建的原始测试向量为 $q_t = (\{d_{t-2}, d_{t-1}, 0, 1, 2\})$. d_{t+1}, d_{t+4} 构建的原始测试向量分别为 $q_{t+1} = (\{d_{t-2}, d_{t-1}, 0, 1, 3\})$, $q_{t+4} = (\{d_{t+2}, d_{t+3}, 0, 1, 2\})$.

将原始训练向量和原始测试向量经过第 2.2 节的方法转变为二维信息矩阵,构成基于时间相关性的训练集和测试集,并使用 CNN 填充模型对缺失数据进行填充,得到时间维度的填充结果。

3.2.2 空间相关性

部署在监测区域内的多个传感器节点所形成的传感器网络,网络中节点与节点间的数据存在空间相关性,因此可以利用与缺失数据在同一传感网络中的其他传感器节点的数据,对缺失数据进行填充。

以传感器节点 n 为例说明基于空间相关的原始训练向量和测试向量的构建,设传感器节点在 k 时刻数据完好,而在 t 时刻数据发生缺失. 因此,在 k 时刻,以其他 $N-1$ 个传感器的数据作为输入,传感器节点 n 的数据作为输出构建原始训练向量,即 $p = (\{d_{1k}, d_{2k}, \dots, d_{n-1,k}, d_{n+1,k}, \dots, d_{Nk}\}, \{d_{nk}\})$. 同理,在 t 时刻的原始测试向量可表示为 $q = (\{d_{1t}, d_{2t}, \dots, d_{n-1,t}, d_{n+1,t}, \dots, d_{Nt}\})$.

同样,将原始训练向量和原始测试向量转变为二维信息矩阵,构成基于空间相关性的训练集和测试集,并使用 CNN 填充模型对缺失数据进行填充,得到空间维度的填充结果。

3.2.3 时空相关性

基于时间相关性的填充是依据缺失数据所属时间序列之间的关系进行填充,得到时间维度的缺失数据估计值;基于空间相关性的填充是利用传感器节点之间的相互关系进行填充,得到空间维度的缺失数据估计值;因此,基于时空相关性的填充方法就是同时利用时间和空间信息对缺失数据进行填充. 本文构建一个 3 层全连接网络整合时间维度和空间维度的估计值,获得时空相关性的填充结果. 基于 CNN 的缺失数据填充方法的基本流程图如图 2 所示。

4 实验与结果分析

4.1 数据源与数据处理

本文采用某大坝测站点的真实数据进行实验,

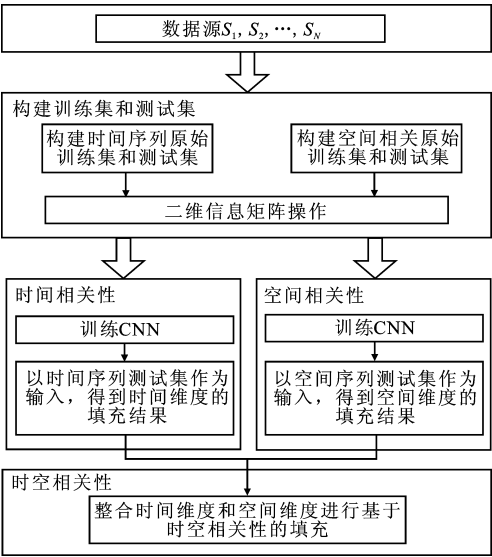


图 2 基于 CNN 的缺失数据填充方法的基本流程图

共有 7 个数据源,每个数据源上记录了近 27 个月的温度数据,其中前 18 个月的监测数据为每天更新一次,后 9 个月的监测数据为每天更新 3 次,将这 7 个数据源的时间序列数据进行时空整合共得到 1 351 个样本数据.由于所得到的温度数据含有少数缺失值,因此在实验过程中将缺失值替换为邻近的前后数据的平均值,形成一个完整的数据集.本文选择传感器节点 n 作为实验对象,随机插入缺失数据,使其拥有 $r\%$ 的缺失数据,其中 $r \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$.

4.2 建立卷积神经网络填充模型

本文在 PyCharm 编程环境下使用 Python 和 TensorFlow 机器学习框架建立神经网络模型,以最佳回归效果为目标确立卷积神经网络填充模型结构,模型结构如图 3 所示.

在传感器节点 n 上,基于时间相关性进行填充时获得的原始训练集和测试集有 5 维特征,基于空间相关性进行填充时获得的原始训练集和测试集有 6 维特征.经过二维信息矩阵变换后,最终基于时间相关性的卷积神经网络填充模型的输入数据维度为 $\text{None} \times 5 \times 5$,输出数据维度为 $\text{None} \times 1$;基于空间相关性的卷积神经网络填充模型的输入数据维度为 $\text{None} \times 6 \times 6$,输出数据维度为 $\text{None} \times 1$,其中 None 为模型输入样本个数.

卷积神经网络填充模型,选用大小为 2×2 ,步长为 1 的卷积核进行卷积操作,选用大小为 2×2 ,步长为 2 的池化核进行池化操作,卷积层和池化层的 Padding 都设置为 SAME,激活函数为线性整流

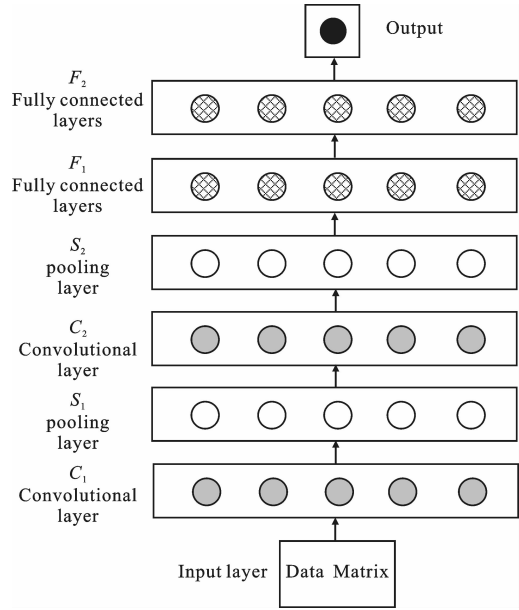


图 3 卷积神经网络填充模型结构

单元 (Rectified Linear Unit, ReLU). 卷积神经网络填充模型以最小化损失函数均方误差 (MSE) 为目标,具体来说就是为了减小填充值与期望的真实值之间的误差.在模型训练初期较大的学习率可加快模型收敛速度,使模型快速靠近真实值;随着训练次数的增加,较小的学习率可以使模型在真实值附近收敛,鉴于以上考虑,模型中的学习率为 $lr = 0.001 \times 0.95^{\text{epoch}}$,其中 0.001 为初始学习率,epoch 为当前迭代次数,模型训练的最大迭代次数 $\text{epoch}_{\max} = 1\,000$.

4.3 结果分析

本文选用希尔不等系数(Theil Inequality Coefficient, TIC)度量填充结果的准确性. TIC 的值介于 0 和 1 之间,度量指标 TIC 的数值越小,说明填充值与真实值的误差越小,回归拟合度越好,填充精度越高. TIC 公式为:

$$\text{TIC} = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n} \sum_{i=1}^n \hat{y}_i^2} + \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2}} \quad (5)$$

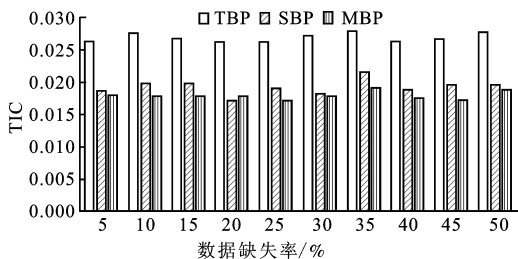
式中, $\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$ 表示均方根误差 (Root Mean Square Error, RMSE); y_i 表示真实值, $\hat{y}_i$ 表示填充值, n 为缺失数据个数.

本文采用三层 BP 神经网络作为基本的缺失数据填充模型,与本文提出的基于卷积神经网络的缺失数据填充模型进行对比.依据填充时考虑到的时间与空间两方面的相关性,分别利用 MBP, TBP,

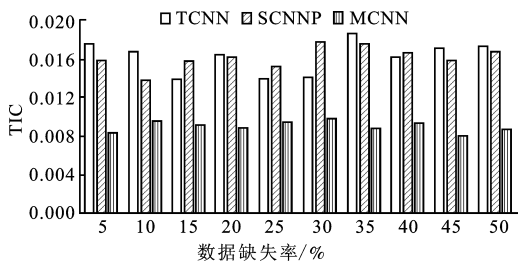
SBP 表示考虑时空多维度相关性、只考虑时间相关性和只考虑空间相关性的 BP 填充方法. 同理, 利用 MCNN, TCNN, SCNN 表示考虑时空多维度、时间和空间相关性的 CNN 填充方法. 本文选用十折交叉验证来确定模型的最优参数, 在实验过程中, 分别对每一个缺失率的实验重复 20 次, 取平均值作为最终的实验结果.

4.3.1 单维度和多维度相关性对填充结果的影响

根据第 3 节的介绍可知, 本文利用传感器网络中节点自身的时间序列数据和节点间的空间关系对缺失数据进行填充, 进而考虑利用多维度相关性对缺失数据进行填充. 图 4 给出利用单维度和多维度相关性对缺失数据进行填充的结果.



(a) BP 填充



(b) CNN 填充

图 4 单维度 vs. 多维度相关性对 TIC 的影响

从图中可以看出, 利用时空多维度相关性的缺失数据填充结果普遍优于单维度相关性的缺失数据填充结果. 这是由于利用时空多维度相关性的填充结果结合了时间和空间两个维度数据的变化规律, 能够对已有数据进行更大程度的挖掘, 进而填充结果的准确性更高.

4.3.2 多维度相关性填充结果对比

本节将给出考虑时空多维度相关性情况下基于 CNN 的缺失数据填充方法和基于 BP 的缺失数据填充方法在评价指标 TIC 上的比较结果. 由图 5 可以看出, 在相同缺失率和相同的数据变换方式下, 与基于 BP 的填充方法相比, 基于 CNN 的填充方法有更好的填充效果. 虽然 BP 神经网络与卷积神经网络具有相似的训练方式, 但 CNN 具有的独特网络

结构能够隐性地学习数据的抽象特征, 并完成对抽象特征地提取和回归操作, 因此 CNN 填充模型有较强的学习能力, 能够获得更好的填充结果.

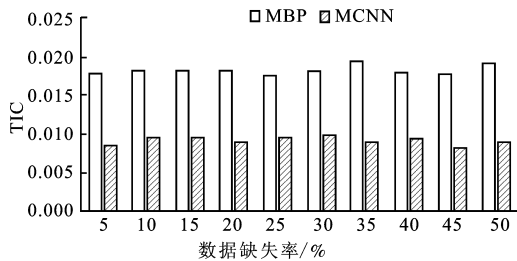


图 5 填充结果对比

5 结束语

本文分别研究传感器网络的数据特点和卷积神经网络技术, 在此基础上提出一种基于卷积神经网络的缺失数据填充方法. 首先基于时间相关性和空间相关性, 使用 CNN 填充模型分别对缺失数据数据进行单维度相关性的填充. 然后考虑数据时空相关性, 得到基于时空相关性的填充方法 MCNN. 最后在真实数据集上进行实验验证, 实验结果表明基于时空多维度相关性的填充结果好于单维度相关性的填充结果, 具有较高的填充精度和回归拟合度. 同时, 与 BP 神经网络相比, 本文所提方法的填充准确性更高. 下一步, 将考虑时间、空间和属性三个维度的相关性, 使用神经网络模型对缺失数据进行填充.

参考文献:

- [1] 李建中, 王宏志, 高宏. 大数据可用性的研究进展[J]. 软件学报, 2016, 27(7): 1605-1625.
- [2] 陈志奎, 吕爱玲, 张清辰. 基于属性重要性的不完备数据填充算法[J]. 微电子学与计算机, 2013, 30(7): 167-172.
- [3] CHANG C Y, WANG C R, LEE S J. Novel imputation for time series data[C]//Machine Learning and Cybernetics (ICMLC), 2015 International Conference on. Guangzhou, China, IEEE, 2015: 916-920.
- [4] ZHAO N, LI Z, LI Y. Improving the traffic data imputation accuracy using temporal and spatial information[C]//Intelligent Computation Technology and Automation (ICICTA), 2014 7th International Conference on. Changsha, China, IEEE, 2014: 312-317.
- [5] 卜范玉, 陈志奎, 张清辰. 基于深度学习的完整大数据填充算法[J]. 微电子学与计算机, 2014, 31(12): 173-177.

(下转第 57 页)

5 结束语

针对遗传算法存在的“早熟”问题并进一步加快算法收敛速度、提高收敛精度^[9-10],本文对遗传算法做了两部分的改进:改进交叉和变异概率公式与改进选择策略,通过分别仿真对比两个改进方法,可以看出,通过改进交叉和变异概率公式能够提高收敛精度和速度,提高算法的全局搜索能力,通过改进选择策略可以进一步加快收敛速度.实验证明,本文提出的 MAGA 算法能够加快收敛速度,有效避免局部收敛和提高算法的稳定性和精确度.

参考文献:

[1] 谭丹丹,曹斌,刘俊,等.一种基于配对的改进遗传算法[J]. 计算机应用,2007,12(27):170-171.

[2] 曲志坚,张先伟,曹雁锋,等.基于自适应机制的遗传算法研究[J]. 计算机应用研究,2015,32(11):3222-3225.

[3] 王小平,曹立明.遗传算法:理论、应用与软件实现[M]. 西安:西安交通大学出版社,2002.

[4] SRINIVAS M, PATNAIK L M. Adaptive probabili-

ties of crossover in genetic[J]. IEEE Trans Sys, Man and Cybem, 1994, 24(4): 656-667.

[5] 卢长娜,王如云,陈耀登.自适应遗传算法[J]. 计算机仿真,2006,23(1):172-175.

[6] 邝航宇,金晶,苏勇.自适应遗传算法交叉变异算子的改进[J]. 计算机工程与应用,2006,42(12):93-96.

[7] 张京钊,江涛.改进的自适应遗传算法[J]. 计算机工程与应用,2010,46(11):53-55.

[8] 王慧妮,彭其渊,张晓梅.基于种群相异度的改进遗传算法及应用[J]. 计算机应用,2006,26(3):668-669.

[9] 沐阿华,周绍磊,于晓丽.一种快速自适应遗传算法及其仿真研究[J]. 系统仿真学报,2004,15(1):122-125.

[10] ROYACHKA K, KAROVA M. High-Performance optimization of genetic algorithms [C]//ISSE' 06. 29thInternational Spring Seminar Electronics Technology. Rome, Italy,2006: 395-400.

作者简介:

孔盼盼 女,(1994-),硕士研究生.研究方向为 RFID 性能测试. E-mail:SunLightKong@163.com.

崔英花 女,(1973-),教授,硕士生导师.研究方向为无线射频识别技术.

(上接第 52 页)

[6] 张宏亭,李学仁,孔韬. BP 神经网络在缺失数据估计中的应用[J]. 计算机工程与设计,2007,28(14):3457-3459.

[7] 刘树洁. 基于神经网络风电场测风数据插补与长年代风速推算方法研究[D]. 湖北:武汉大学,2017.

[8] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J]. 计算机学报,2017,40(6):1229-1251.

[9] 王璨,武新慧,李恋卿,等.卷积神经网络用于近红外光谱预测土壤含水率[J]. 光谱学与光谱分析,2018,38(1):36-41.

作者简介:

张网娟 女,(1992-),硕士研究生.研究方向为大数据、数据管理. E-mail: zhang_wang_juan@foxmail.com.

许国艳 女,(1971-),博士,副教授.研究方向为大数据、数据起源、数据管理.

李敏佳 女,(1994-),硕士研究生.研究方向为社交网络、数据管理.

朱 帅 男,(1992-),硕士研究生.研究方向为社会网络、数据管理.