DOI: 10.11992/tis.201807037
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20181224.1044.003.html

缺失数据的混合式重建方法

于本成[1,2]，丁世飞[1]
（1. 中国矿业大学 计算机科学与技术学院，江苏 徐州 221116；2. 徐州工业职业技术学院 信息与电气工程学院，江苏 徐州 221004）

摘　要：缺失数据的问题在各领域中是不可避免的，而传统的数据挖掘算法在处理不完整的数据集时表现不佳。本文将协方差矩阵及协方差矩阵的行列式应用于粒子群优化算法的适应度函数中，并以迭代的方式得出最佳阈值，再使用最佳阈值进行基于进化聚类算法的缺失值重建，解决了阈值的选取困难及其对数据重建结果的影响问题。然后，在自联想极限学习机中调用具有最佳阈值的进化聚类算法，解决了自联想极限学习机输入权值选择的随机性。最后，选取 6 个 UCI 标准数据集及 9 个激活函数来进行验证。实验结果表明，相对于现有的大多数数据重建方法，所提的混合式重建方法可以更有效地完成缺失数据的重建。

关键词：数据挖掘；协方差矩阵；适应度函数；粒子群优化；最佳阈值；进化聚类算法；数据重建；自联想的极限学习机
中图分类号：TP301.6　　**文献标志码**：A　　**文章编号**：1673−4785(2019)05−0947−06

中文引用格式：于本成，丁世飞. 缺失数据的混合式重建方法 [J]. 智能系统学报，2019，14(5)：947–952.
英文引用格式：YU Bencheng, DING Shifei. Hybrid reconstruction method for missing data[J]. CAAI transactions on intelligent systems, 2019, 14(5): 947–952.

Hybrid reconstruction method for missing data

YU Bencheng[1,2], DING Shifei[1]
(1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; 2. School of Information and Electrical Engineering, Xuzhou College of Industrial Technology, Xuzhou 221004, China)

Abstract: The problem of missing data is inevitable in different areas. However, traditional data mining algorithms do not process incomplete data sets well. The covariance matrix and its determinant were applied to the fitness function of particle swarm optimization, and the optimal threshold was obtained through iteration. Then, the missing data were reconstructed based on the evolving clustering method using the optimal threshold, which solved the difficulty in optimal threshold selection and determined its influence on data reconstruction results. Furthermore, the randomness of the auto-associative extreme learning machine was removed by invoking the evolving clustering method with the optimal threshold. Finally, six UCI standard data sets and nine activation functions were selected to verify the method. The results showed that compared with most existing reconstruction methods, the proposed hybrid reconstruction method can complete the reconstruction of the missing data more effectively.

Keywords: data mining; covariance matrix; fitness function; particle swarm optimization; optimal threshold; evolving clustering method; data reconstruction; auto-associative extreme learning machine

鉴于缺失数据重建的重要性，研究人员已经就缺失数据重建问题提出了多种解决方法。粒子群优化 (particle swarm optimization, PSO) 在 1995 年由 Kennedy 和 Eberhart 提出[1-2]。PSO 通过群体内个体之间的信息共享来对问题的解进行协同搜索[3]，即初始化一群随机粒子，并通过迭代找到最优解。在每一次迭代中，粒子通过跟踪局部

收稿日期：2018−07−31.　**网络出版日期**：2018−12−25.
基金项目：国家自然科学基金项目 (61379101).
通信作者：丁世飞. E-mail: dingsf@cumt.edu.cn.

最优值和全局最优值来更新自己的速度与位置[4]。文献 [5] 通过调整惯性权重的取值，提出了自适应混沌粒子群优化算法，该算法避免了粒子早熟收敛情况。文献 [6] 中 Krishna 和 Ravi 提出了一种基于粒子群优化和矩阵协方差结构的数据重建方法，他们使用 PSO 重建缺失值。

进化聚类算法 (evolving clustering method, ECM) 是一步到位的快速聚类算法，在 ECM 中，由用户定义的阈值参数 τ 会影响群集合数量的估计，τ 值太大或太小都不利于找出群集合数量[7-8]。Ravi 等[9-10] 提出了 4 种用于重建的混合方法。在线重建中使用了具有广义回归神经网络的 ECM(ECM+GRNN)，在离线重建中使用了 K-means+GRNN 和 K-medoids+GRNN 以及具有多层感知机的 K-medoids(K-medoids+MLP)。他们虽然提出了基于 ECM 的数据重建，但 τ 值选择涉及了试错法，结果都不同程度地受到 τ 值的影响。

极限学习机 (extreme learning machine, ELM) 是由 Huang 等[11-12] 提出的，它是一种新颖的前馈神经网络，不需要权重更新。目前 ELM 的理论与算法研究主要集中在随机生成参数的优化、最优外权的求解、最优隐藏层节点个数的选取、ELM 核函数、在线极限学习机算法等方面[13]。文献 [14] 发现自联想的极限学习机 (auto associative extreme learning machine, AAELM) 在同一个数据集合中的不同运行产生了不同的结果。有时，连接输入层和隐藏层的随机加权会使结果出现很大的波动。在文献 [15] 中 Ravi 和 Krishna 为重建提出了多种在线和离线方法，即粒子群优化训练后的自动关联神经网络 (PSOAANN)、粒子群优化训练后的自动关联子波神经网络 (PSOAAWNN)、径向基函数自动关联神经网络 (RBFAANN)、广义回归自动关联神经网络 (GRAANN)，这些算法仍有待于进一步改进。

鉴于以上研究中存在的 PSO 重建缺失值效率低、τ 值的选取会影响基于 ECM 的数据重建、AAELM 结构中连接输入层和隐藏层的随机加权导致重建结果波动较大以及文献 [15] 中所提的神经网络的重建效率低等问题，本文根据协方差矩阵具有旋转不变性的特征[16]，在 PSO 适应度函数中用到了协方差矩阵及协方差矩阵的行列式，选取了最佳 τ 值，使用具有最佳 τ 值的 ECM 进行缺失数据重建，我们称之为 PSOECM 方法，PSOECM 方法解决了 τ 值的选取困难及其对基于 ECM 重建结果的影响问题。随后，在 AAELM中调用具有最佳 τ 值的 ECM，去除了 AAELM 的随机性，我们称之为改进型的 AAELM(MAAELM)。

1　基础理论

1.1　PSO 算法

在 D 维搜索空间中，由 n 个粒子组成的群体中粒子 $i(i=1,2,\cdots,n)$ 的位置表示为 D 维位置矢量 $\boldsymbol{z}_i=(z_{i1},z_{i2},\cdots,z_{id},\cdots,z_{iD})$，每次迭代中粒子 i 移动的距离为速度矢量或飞行速度 $\boldsymbol{v}_i=(v_{i1},v_{i2},\cdots,v_{id},\cdots,v_{iD})$，粒子迄今为止搜索到的最优位置 $\boldsymbol{p}_i=(p_{i1},p_{i2},\cdots,p_{id},\cdots,p_{iD})$，整个粒子群迄今为止搜索到的最优位置为 $\boldsymbol{p}_g=(p_{g1},p_{g2},\cdots,p_{gd},\cdots,p_{gD})$，每次迭代中任一粒子根据式 (1)、式 (2) 来更新自己的速度和位置：

$$v_{id}^{k+1}=\omega v_{id}^{k}+a_1r_1\left(p_{id}-z_{id}^{k}\right)+a_2r_2\left(p_{gd}-z_{id}^{k}\right) \tag{1}$$

$$z_{id}^{k+1}=z_{id}^{k}+v_{id}^{k+1} \tag{2}$$

式中：$i=1,2,\cdots,n$；ω 是惯性权重；k 是迭代次数；a_1、a_2 是加速系数；r_1、r_2 是 [0,1] 范围内的随机数。PSO 算法的描述如下：

1) 随机初始化群体，设定粒子的位置和速度；

2) 根据适应度函数计算粒子的适应度值，选取具有最优适应度值的粒子位置作为 $\boldsymbol{p}_g$，每个粒子当前位置为 $\boldsymbol{p}_i$；

3) 根据式 (1)、式 (2) 更新粒子的速度和位置；

4) 把每个粒子的适应度值与 $\boldsymbol{p}_i$ 的适应度值进行比较，若优于 $\boldsymbol{p}_i$ 的值，则将其值设为 $\boldsymbol{p}_i$；

5) 把每个粒子的适应度值与 $\boldsymbol{p}_g$ 的适应度值进行比较，若优于 $\boldsymbol{p}_g$ 的值，则将其值设置为 $\boldsymbol{p}_g$；

6) 检查是否满足终止条件，如果满足则终止迭代，否则返回 2)。

1.2　ECM 算法

创建新群集 C 时，定义群集中心 C'，并将群集半径 R 初始为零。随着样本的相继出现，已经创建的群集可以通过改变群集中心 C' 位置和增加群集半径 R 来更新。当群集半径 R 达到阈值 τ 值时，将不再更新群集。ECM 算法过程如下：

1) 创建第一个群集 C_1，并将输入数据中的第一个样本作为群集中心 C'_1，设置群集半径 $R_1=0$。

2) 如果输入数据流的所有样本都已处理完毕，则算法结束。否则，取当前样本 x_i，计算 x_i 与已经创建的所有 n 个集群中心 C'_j 之间的距离，$D_{ij}=\|x_i-C'_j\|$，其中 $j=1,2,\cdots,n$。

3) 如果存在群集中心 C'_j，其中 $j=1,2,\cdots,n$，使得 $D_{ij}=\|x_i-C'_j\|\leqslant R$，则假定当前样本 x_i 属于最近群集 C_m，$D_{im}=\|x_i-C'_m\|=\min\left(\|x_i-C'_j\|\right)$。在这种情况下，既不创建新群集，也不更新现有群集，并

返回到 2), 否则进入 4)。

4) 从已经创建的所有 n 个集群的中心中, 通过计算 $S_{ij}=D_{ij}+R_j, j=1,2,\cdots,n$, 找出一个群集 C_a。再通过计算算出最小的 S_{ia} 值, $S_{ia}=D_{ia}+R_a=\min\{S_{ij}\}$, $j=1,2,\cdots,n$, 来找出 C_a 的群集中心 C'_a。

5) 如果 $S_{ia}>2\tau$, 则样本 x_i 不属于任何现有群集, 那么以与 1) 的相同方式创建新集群, 执行 2)。

6) 如果 $S_{ia}\leqslant 2\tau$, 则通过移动群集中心 C'_a 和增加群集半径 R_a 来更新群集 C_a, 返回 2)。

ECM 算法不保持已传递样本的任何信息, 但任一群集 C_i 的群集中心 C'_i 到该群集的最远样本之间距离都小于阈值 τ, 即 $\max(R_i)<\tau$。

在 ECM 算法中, 向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间的距离计算使用归一化欧几里德距离, 即

$$\|\boldsymbol{x}-\boldsymbol{y}\|=\left(\sum_{i=1}^{q}|\boldsymbol{x}_i-\boldsymbol{y}_i|^2\right)^{1/2}/q^{1/2}\ ,\ \boldsymbol{x},\boldsymbol{y}\in\mathbf{R}^q \tag{3}$$

在 5)、6) 中 τ 值的大小影响到群集合数量, 所以 τ 值的选取影响到了基于 ECM 的数据重建结果。

1.3　ELM 算法

输入层的节点个数为 n, 隐藏层节点个数为 L, 输出层节点个数为 m, a_{ij} 代表第 i 个输入层节点与第 j 个隐藏层节点间的权值, b_j 代表隐藏层中第 j 个节点的偏置。β_{jk} 是需要计算的值, 代表第 j 个隐藏层节点与第 k 个输出层节点间的权值。训练集实例个数为 N 的输入矩阵 $\boldsymbol{X}$ 以及输出矩阵 $\boldsymbol{T}$ 分别为

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{bmatrix} \tag{4}$$

$$\boldsymbol{T}=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1N} \\ t_{21} & t_{22} & \cdots & t_{2N} \\ \vdots & \vdots & & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nN} \end{bmatrix} \tag{5}$$

第 i 个实例在第 j 个隐藏层神经元上的输出为 $G(a_i,b_i,x_i)$, 整个的输出层值为

$$\sum_{j=1}^{L}\beta_j G(a_j,b_j,x_i)=t_i,\quad i=1,2,\cdots,N \tag{6}$$

式 (6) 也可以表示为

$$\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{T} \tag{7}$$

式中 $\boldsymbol{H}$ 表示隐藏层的矩阵。$\boldsymbol{H}$ 矩阵第 i 行代表输入层中第 i 个实例在隐藏层所有神经元上的输出, $\boldsymbol{H}$ 矩阵的第 j 列代表所有训练样本在第 j 个隐藏层神经元上的输出, 即

$$\boldsymbol{H}(a_1,\cdots,a_L,b_1,\cdots,b_l,x_1,\cdots,x_n)=\begin{bmatrix} G(a_1,b_1,x_1) & \cdots & G(a_L,b_L,x_1) \\ \vdots & & \vdots \\ G(a_1,b_1,x_N) & \cdots & G(a_L,b_L,x_N) \end{bmatrix}_{N\times L} \tag{8}$$

$$\boldsymbol{\beta}=\begin{bmatrix} \boldsymbol{\beta}_1^{\mathrm{T}} \\ \boldsymbol{\beta}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\beta}_L^{\mathrm{T}} \end{bmatrix}_{L\times m}\quad \boldsymbol{T}=\begin{bmatrix} \boldsymbol{t}_1^{\mathrm{T}} \\ \boldsymbol{t}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{t}_N^{\mathrm{T}} \end{bmatrix}_{N\times m} \tag{9}$$

在已知权值和偏置的情况下, 上面问题的求解就转化为求解线性系统 $\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{T}$ 的最小范数最小二乘解:

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{\dagger}\boldsymbol{T} \tag{10}$$

式中: $\boldsymbol{H}^{\dagger}$ 是 $\boldsymbol{H}$ 的 Moore-penrose 广义逆矩阵; $\hat{\boldsymbol{\beta}}$ 的范数是最小且唯一的。

2　提出的混合式重建方法

2.1　PSOECM 方法

全部数据记录 $\boldsymbol{X}_t$ 可以分为两个部分: 用于训练模型的完整记录集 $\boldsymbol{X}_c$ 和用于检验模型的不完整记录集 $\boldsymbol{X}_{ic}$。

PSOECM 方法步骤:

1) 计算出 $\boldsymbol{X}_c$ 的协方差矩阵。

2) 在具有 PSO 随机初始化 τ 值的 $\boldsymbol{X}_c$ 上应用 ECM。

3) 对 $\boldsymbol{X}_{ic}$ 执行基于 ECM 的重建: 通过测量除去缺失值的不完整记录与除去相同位置上值的群集中心 C' 之间的欧几里德距离确定最近群集中心, 由最近群集中心的对应属性值重建不完整记录的属性值 (x_k)。欧几里德距离的测定公式为

$$D_j=\sum_{i=1;i\neq k}^{n}\left|x_i-C'_j\right|^2 \tag{11}$$

式中: j 为群集中心的数量; n 为每条记录中完整成分的数量。

4) 数据重建后计算 $\boldsymbol{X}_t$ 的协方差矩阵。如果 $\boldsymbol{X}_t$ 为 $(m\times n)$ 秩序的矩阵, 则它的协方差矩阵 $\boldsymbol{T}_{\mathrm{COV}}$ 就是一个 $n\times n$ 矩阵。如果 $\mathrm{MSE}(\boldsymbol{X}_{\mathrm{COV}},\boldsymbol{T}_{\mathrm{COV}})<\varepsilon$ 且 $(|\mathrm{Det}(\boldsymbol{X}_{\mathrm{COV}})-\mathrm{Det}(\boldsymbol{T}_{\mathrm{COV}})|)<\varepsilon$, 则退出计算。否则, 调用 PSO 选出改善后的 τ 值。其中 ε 为预先设定的小正值, $\mathrm{MSE}(\boldsymbol{X}_{\mathrm{COV}},\boldsymbol{T}_{\mathrm{COV}})$ 为 $\boldsymbol{X}_{\mathrm{COV}}$ 和 $\boldsymbol{T}_{\mathrm{COV}}$ 元素之间的均方差, $\mathrm{Det}(\boldsymbol{X}_{\mathrm{COV}})$ 是 $\boldsymbol{X}_{\mathrm{COV}}$ 的行列式, $\mathrm{Det}(\boldsymbol{T}_{\mathrm{COV}})$ 是 $\boldsymbol{T}_{\mathrm{COV}}$ 的行列式。

5) 重复 1) ~ 4) 直至收敛。

计算平均绝对百分比误差 (mean absolute percentage error, MAPE) 值:

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i}\right|\times 100\% \tag{12}$$

式中：x_i 为实际值；$\hat{x}_i$ 为预测值；n 为缺失值的全部样本数量。

PSOECM 方法采用与文献 [6] 相同的适应度函数 $\mathrm{MSE}(\boldsymbol{X}_{\mathrm{COV}},\boldsymbol{T}_{\mathrm{COV}})$ 和 $(|\mathrm{Det}(\boldsymbol{X}_{\mathrm{COV}})-\mathrm{Det}(\boldsymbol{T}_{\mathrm{COV}})|)$，但文献 [6] 使用 PSO 重建缺失值，而 PSOECM 方法使用 PSO 以迭代的方式完成了上述两个适应度函数的最小化工作，只有两个适应度函数在两个连续迭代中都小于预先设定 ε 值才停止运算，并计算出最佳 τ 值，再在 ECM 中使用 PSO 选择最佳 τ 值进行缺失数据重建。这样不仅可以得出最佳的数据重建，还可以保存数据的协方差结构。

2.2　MAAELM 方法

MAAELM 方法采用 PSOECM 与 AAELM 混合重建缺失数据。MAAELM 结构如图 1 所示。

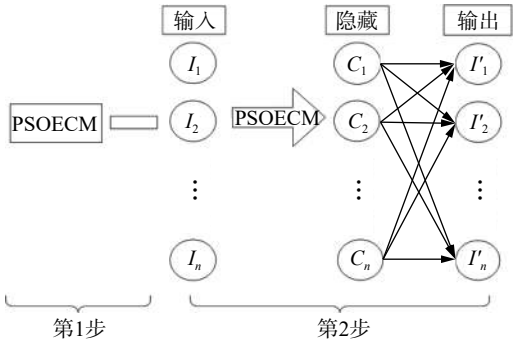

图 1　MAAELM 结构

Fig. 1　Architecture of the MAAELM

MAAELM 方法步骤：

1) 将数据归一化至 [0,1] 范围内。

2) 将数据集合分为完整记录集合和不完整记录集合。

3) 在 1) 中执行基于 PSOECM 的重建，确定群集中心。

4) 在 2) 中使用 1) 中得出的最佳 τ 值在完整记录集合中应用 ECM。这相当于使用 1) 中得到的群集中心作为 MAAELM 结构中的隐藏节点。

5) 执行 PSOECM 方法的 3)。

6) 计算得到各个群集中心之间的归一化欧几里德距离。

为了估算出隐藏层和输出层之间的权重，在 6) 得到的距离中应用激活函数并进行非线性转换，再应用 Moore-Penrose 广义逆矩阵得出 $\boldsymbol{H}$。

最后，根据文献 [12] 使用 Moore-Penrose 广义逆矩阵求解 $\boldsymbol{H\beta}=\boldsymbol{T}$ 估算出隐藏层和输出层之间的权重，其中 $\boldsymbol{\beta}$ 为权向量，$\boldsymbol{T}$ 为目标向量。利用式 (12) 计算平均绝对百分误差 (MAPE) 值。

3　选取实验数据集与激活函数

实验选取 UCI 机器学习数据库中的 6 个标准数据集来进行验证，实验数据集如表 1 所示。同时，在选取的实验数据集上使用 9 个激活函数来研究它们对文章所提方法的影响。实验选取激活函数如表 2 所示。所选数据集中除 Auto-mpg 中的马力属性值存在缺失，其他 5 个数据集均不存在属性缺失值，所以通过随机删除原始数据集的一些值来进行实验，并创建了除目标变量以外的所有变量中的缺失值。每一个数据集被分成 10 个相等的小集合，其中 9 个小集合经过聚类处理，剩下的 1 个留下为缺失值备用。

为了在每一个小集合中创建缺失值，随机删除了近 10% 的值 (单元)，并确保从每个记录中删除至少一个单元。因此，在 10 倍交叉验证中，有不同缺失记录的 10 个小集合。

对于完整记录集合中的各个小集合，将它们从全部记录中分理并用于聚类。在完整记录集合中应用 ECM 算法，并通过最近群集中心属性的对应值重建出不完整记录集合中的属性缺失值。

使用 PSO 优化算法和文献 [6] 提及的两个适应度函数为 PSOECM 选出最佳 τ 值，并将相同的 τ 值提供给 MAAELM。对于所有数据集合，对比了本文所提方法与文献 [6, 9-10, 15, 17] 所提多种混合方法的 MAPE 平均值。

表 1　实验数据集

Table 1　Data sets for the experiment

数据集名称	实例数	属性数	链接地址
Auto-mpg	398	9	HTTP://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg
Boston Housing	506	13	HTTP://archive.ics.uci.edu/ml/machine-learning-databases/housing
Forest Fires	517	12	HTTP://archive.ics.uci.eduml/machine-learning-databases/forest-fires
Iris Plants	150	4	HTTP://archive.ics.uci.edu/ml/machine-learning-databases/iris
Spectf heart	267	45	HTTP://archive.ics.uci.edu/ml/machine-learning-databases/spect
Wine ecognition	178	13	HTTP://archive.ics.uci.edu/ml/machine-learning-databases/wine

表 2　激活函数表

Table 2　Activation functions

函数名称	函数公式	函数名称	函数公式
Sin	$H=\sin(x)$	Tribas	$\mathrm{Tribas}(x)=\begin{cases}1-\mathrm{abs}(x), & -1\leqslant x\leqslant 1\\ 0, & \text{其他}\end{cases}$
Sinh	$\mathrm{Sinh}(x)=\dfrac{e^{x}-e^{-x}}{2}$	Radial basis function	$\mathrm{RBF}(x)=\exp\left(-\dfrac{x^{2}}{2\sigma^{2}}\right)$, σ 为宽度
Sigmoid	$\mathrm{Sigmoid}(x)=\dfrac{1}{1+e^{-x}}$	Softplus	$\mathrm{Softplus}(x)=\log(1+e^{x})$
Bipolar sigmoid	$\mathrm{Bsigmoid}(x)=\dfrac{1-e^{-x}}{1+e^{-x}}$	Gaussian	$\mathrm{Gaussian}(x)=a\exp\dfrac{-(x-b)^{2}}{c^{2}}$, a、b、c是实数
Hardlim	$\mathrm{Hardlim}(x)=\begin{cases}1, & x\geqslant 0\\ 0, & \text{其他}\end{cases}$		

4　实验结果和分析

不同激活函数作用于 MAAELM 所得的 MAPE 值以及 PSOECM、MAAELM 与其他算法比较的结果如图 2 和图 3 所示。

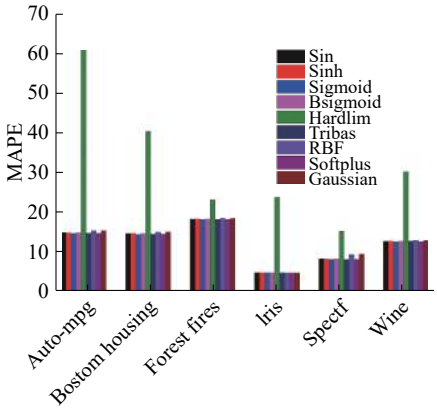

图 2　不同激活函数对 MAAELM 的影响

Fig. 2　Influence of different activation functions on the MAAELM

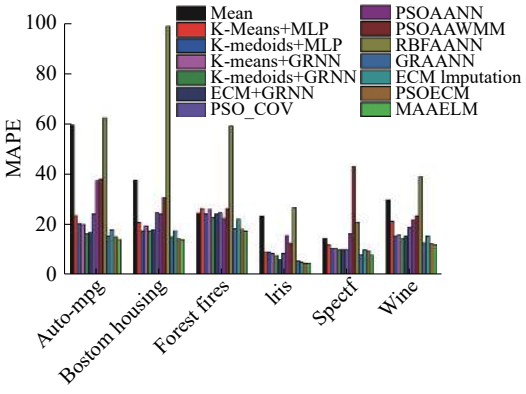

图 3　不同算法的 MAPE 值

Fig. 3　MAPE value of different algorithms

根据图 2 所展示的不同激活函数作用于 MAAELM 所得的 MAPE 值可以发现：Sigmoid 在所有激活函数中的表现最佳，Hardlim 激活函数表现最差，而其他激活函数对于 MAAELM 的 MAPE 值影响基本相同。Hardlim 激活函数表现最差是因为它将一个输入空间只分割为 0 和 1 两个类别。

图 3 中将本文所提算法与 Krishna M 和 Ravi V[6] 的 PSO_COV 算法，Nishanth 和 Ravi[9] 的 K-means+GRNN、K-medoids+MLP、K-medoids+GRNN、ECM+GRNN 等算法，Gautam 和 Ravi[10] 的 ECM Imputation 算法，Ravi 和 Krishna[15] 的 PSOAANN、PSOAAWNN、RBFAANN、GRAANN 等算法，Ankaiah 和 Ravi[17] 的 K-Means+MLP 算法的结果进行对比，对比结果显示了最佳 τ 值在所提方法中可以更有效地进行基于 ECM 的重建，以及在大部分数据集合上局部学习和整体学习混合使用优于文献 [6, 9-10, 15, 17] 所提方法。

在 Auto-mpg 数据集合方面，只有 K-medoids+GRNN、ECM+GRNN 和 GRAANN 这 3 种混合法的结果与 PSOECM 方法接近，分别落后 1.31%、1.65% 和 0.19%。PSOECM 通过选择最佳 τ 值，在 Auto-mpg 数据集合中的表现优于 ECM 重建。将 PSOECM 得出的相同 τ 值带入 MAAELM 时，误差又降低了 0.96%。

在 Boston Housing 数据集合方面，除了 GRAANN 方法与 PSOECM 方法相差 0.88% 之外，其他方法的 MAPE 值至少比 PSOECM 高 3%。PSOECM 通过选择最佳 τ 值，在 Boston Housing 数据集合中的表现同样优于 ECM 重建。在 MAAELM 中应用 PSOECM 得出的最佳 τ 值之后，MAPE 值便可以进一步降低 0.32%。

在 Forest fires 数据集合方面，可以观察到与 Boston Housing 数据集合相似的性能。除了 GRAANN 落后 PSOECM 的结果 0.13% 之外，其他方法的 MAPE 值比 PSOECM 至少高 4%。PSOECM 通过选择最佳 τ 值，MAPE 同样有所下降。在 MAAELM

中应用 PSOECM 得出最佳 τ 值之后，误差又降低了 0.68%。

除了在 Spectf 数据集合中，PSOECM 略逊于 GRAANN 之外，在 Iris、Spectf 和 Wine recognition 数据集合中，PSOECM 与 MAAELM 同样表现出了类似在 Auto-mpg、Boston Housing、Forest fires 数据集合中的优势。

经上述实验结果的分析得出：1)PSOECM 通过选择最佳 τ 值，在各个数据集合中的表现优于 ECM 重建；2) 将 PSOECM 得出的相同 τ 值代入 MAAELM 时，所得 MAPE 值均有所降低。

5　结束语

本文提出了 2 种新颖的缺失数据的混合式重建方法，并使用 6 个数据集验证了所提方法的有效性。发现由 PSO 为 ECM 选出的最佳 τ 值在 PSOECM 和 MAAELM 的优异性能方面起到了重要作用，解决了 τ 值的选取困难和 τ 值对 ECM 重建结果的影响问题，同时去除了 AAELM 的随机性。下一步研究将增大实验数据集，验证本文所提方法在原始数据缺失不同百分比时的结果，以及使用更多的激活函数来进一步验证所提方法的有效性，并对所提方法与现有方法进行威尔克森符号秩检验，验证所提方法的显著性。

参考文献：

[1] KENNEDY J. Particle swarm optimization[M]//SAMMUT C, WEBB G I. Encyclopedia of Machine Learning. Boston, MA: Springer, 2010.

[2] EBERHART R C, SHI Y. Comparing inertia weights and constriction factors in particle swarm optimization[C]// Proceedings of the 2000 Congress on Evolutionary Computation. La Jolla, USA, 2000: 84−88.

[3] 张庆科. 粒子群优化算法及差分进行算法研究 [D]. 济南: 山东大学, 2017.
ZHANG Qingke. Research on the particle swarm optimization and differential evolution algorithms[D]. Ji'nan: Shandong University, 2017.

[4] 王永贵, 林琳, 刘宪国. 基于改进粒子群优化的文本聚类算法研究 [J]. 计算机工程, 2014, 40(11): 172–177.
WANG Yonggui, LIN Lin, LIU Xianguo. Research on text clustering algorithm based on improved particle swarm optimization[J]. Computer engineering, 2014, 40(11): 172–177.

[5] 徐林. 粒子群优化算法的改进及其应用研究 [J]. 西安文理学院学报 (自然科学版), 2017, 20(4): 51–54.
XU Lin. Research on improvement and application of the particle swarm optimization algorithm[J]. Journal of Xi'an University (natural science edition), 2017, 20(4): 51–54.

[6] KRISHNA M, RAVI V. Particle swarm optimization and covariance matrix based data imputation[C]//Proceedings of 2013 IEEE International Conference on Computational Intelligence and Computing Research. Enathi, India, 2013: 1–6.

[7] KASABOV N K, SONG Qun. DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction[J]. IEEE transactions on fuzzy systems, 2002, 10(2): 144–154.

[8] KASABOV N, SONG Qun, MA Tianmin. Fuzzy-neuro systems for local and personalized modelling[M]// NIKRAVESH M, KACPRZYK J, ZADEH L A. Forging New Frontiers: Fuzzy Pioneers II. Berlin, Heidelberg: Springer, 2008: 175−197.

[9] NISHANTH K J, RAVI V. A computational intelligence based online data imputation method: an application for banking[J]. Journal of information processing systems, 2013, 9(9): 633–650.

[10] GAUTAM C, RAVI V. Evolving clustering based data imputation[C]//Proceedings of 2014 International Conference on Circuits, Power and Computing Technologies. Nagercoil, Tamil Nadu, India, 2014: 1763–1769.

[11] HUANG Guangbin, ZHU Qinyu, SIEW C K. Extreme learning machine: a new learning scheme of feedforward neural networks[C]//Proceedings of 2004 IEEE International Joint Conference on Neural Networks. Budapest, Hungary, 2004: 985–990.

[12] HUANG Guangbin, ZHU Qinyu, SIEW C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1/2/3): 489–501.

[13] 任阳晖. 极限学习机算法及应用研究 [D]. 沈阳: 沈阳航空航天大学, 2017.
REN Yanghui. Extreme learning machine alorithm and application[D]. Shenyang: Shenyang Aerospace University, 2017.

[14] GAUTAM C, RAVI V. Data imputation via evolutionary computation, clustering and a neural network[J]. Neurocomputing, 2015, 156: 134–142.

[15] RAVI V, KRISHNA M. A new online data imputation method based on general regression auto associative neural network[J]. Neurocomputing, 2014, 138: 106–113.

[16] 申小征. 基于维数约简的区域协方差矩阵及其在人脸识别中的应用 [D]. 云南: 云南财经大学, 2017.

[17] ANKAIAH N, RAVI V. A novel soft computing hybrid for data imputation[C]//Proceedings of the 7th International Conference on Data Mining. Las Vegas, Nevada, USA, 2011.

作者简介：

于本成，男，1981 年生，副教授，博士，主要研究方向为人工智能与数据挖掘。参与国家、省级科研课题 2 项，授权专利、软件著作权 22 项。发表学术论文 20 余篇。

丁世飞，男，1963 年生，教授，博士生导师，CCF 理事，CAAI 理事，主要研究方向为人工智能与模式识别。主持国家、省级课题 8 项，取得发明专利 10 项。发表学术论文 200 余篇，出版专著 4 部。