降噪分层映射算法在多维聚类分析中的优化研究

刘　云，张　轶，郑文凤
(昆明理工大学信息工程与自动化学院，昆明 650500)

摘　要：为了在多维聚类分析中运用有效的深度特征选择方法排除冗余和无关的特征属性，学习数据元素的非线性关系提取最佳特征，提出一种降噪分层映射算法(DHM). 首先，基于降噪自动编码器构建非循环神经网络，容错数据经过隐藏层加权和激活函数的训练获取输入数据的非线性关系得到特征空间，实现特征重构选取最佳特征. 其次，特征空间用于调整自组织特征映射神经网，通过计算最小化加权平方欧式距离寻找匹配的获胜神经元. 最后，结合特征选择网络和无监督聚类网络为降噪分层映射神经网，通过整体模型迭代训练，使权重参数和偏差向量同时得到优化，实现有效的无监督聚类方案. 在真实数据集上的实验结果表明，同 AESOM，DCSOM 和 S-SOM 算法相比，DHM 算法在提高聚类质量及准确性方面有更好的表现.

关键词：特征选择；无监督聚类；降噪自动编码器；自组织特征映射
中图分类号：TP3-05　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2022.013001

Optimization research of denoised hierarchical mapping analysis for multidimensional cluster analysis

LIU Yun, *ZHANG Yi*, *ZHENG Wen-Feng*
(Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China)

Abstract: A de-noising hierarchical mapping (DHM) algorithm is proposed in order to use effective deep feature selection methods in multi-dimensional clustering analysis to eliminate redundant and irrelevant features and learn the nonlinear relationship of data elements to extract the best features. In the algorithm, an acyclic neural network is first built based on the denoising autoencoder. Specifically, the fault-tolerant data are trained by hidden layer weighting and activation function to obtain the nonlinear relationship of the input data and the feature space. The features are reconstructed and the best features are selected. Secondly, the feature space is used to adjust the self-organizing feature map neural network. the minimized weighted squared Euclidean distance is calculated to find the matching winning neuron. Finally, the feature selection network and the unsupervised clustering network are combined to construct the noise reduction hierarchical map neural network. The noise reduction hierarchical map neural network is iteratively trained and the weight parameter and the deviation vector are optimized at the same time to realize an effective unsupervised clustering scheme. Experimental results on real data sets show that, compared with AESOM, DCSOM and S-SOM algorithms, the DHM algorithm has better performance in the quality and accuracy of clustering.

Keywords: Feature selection; Unsupervised clustering; Denoising autoencoder; Self-organizing feature map

收稿日期：2020-12-02
基金项目：国家自然科学基金(61761025)；云南省重大科技专项计划(202002AD080002)
作者简介：刘云(1973—)，男，云南昆明人，副教授，主要从事数据挖掘、数据分析、区块链等研究. E-mail：liuyun@kmust.edu.cn
通讯作者：张轶. E-mail：1422261052@qq.com

1 引　言

聚类分析基于数据元素的邻近度进行类分离，根据集群内部的特征同质性将数据划分为多个子集. 多维数据因其存在大量元素影响数据特征的表示，使得聚类算法的预测性能随特征复杂性的增加而降低. 最新的研究表明，在多维聚类算法中引入不同的神经网络模型代替传统的初始化方案，可以分析深层特征之间的关联关系提高学习能力，提取输入数据的最佳特征属性来实现无监督的聚类任务[1,2].

Guo 等[3]提出一种自编码自组织特征映射 AESOM (Autoencoder-Self Organizing Mapping)算法，通过将两种非线性的降维技术集成到混合无监督的深度学习架构中，从数据中提取高度相关的低维特征，算法可以检测并减少数据噪声和缺陷的负面影响. Aly 等[4]提出深度卷积自组织特征映射 DCSOM (Deep Convolutional Self-Organizing Map)算法，将多个卷积自组织特征映射 SOM (Self-Organizing Feature Map)层级联以创建深度网络结构，对最终卷积 SOM 层产生的特征空间进行计算获得有效的特征表示，但在算法中未实施适当的停止标准可能会增加计算成本. 文献[5]提出平滑自组织图 S-SOM(Smoothed Self-Organizing Map)算法，基于输入向量与其最接近的码本向量之间的互补指数距离提取有效特征，减少嘈杂数据中异常特征的影响提高聚类性能，但并未考虑权重对指数距离的影响.

为了从多维数据的非线性深层结构中选取有效的特征属性进行无监督聚类分析[6]，本文提出一种降噪分层映射 DHM (Denoised Hierarchical Mapping)算法. 首先，基于降噪自动编码器 DAE (Denoising Autoencoder)的非循环神经网络结构结合权重和偏置参数得到激活函数，通过激活函数的训练实现特征重构获取特征空间[7]. 其次，采用自组织特征映射神经网替换 DEA 的输出层，依据训练得到的特征空间来调整 SOM，通过计算最小加权平方欧式距离来寻找获胜神经元. 最后，迭代学习训练整体调节 DHM 算法的权重系数和偏置参数，实现有效的无监督聚类方案. 仿真结果表明，对比同类算法，DHM 算法在聚类准确性和鲁棒性方面有所改善.

2 降噪分层映射(DHM)模型

聚类分析的性能取决于选取数据特征的有效性，研究重点是采用有效的方法分析数据对象之间的离散性或相异性信息，选取输入数据属性表示的最佳特征用于数据分类.

2.1 DHM 算法模型

良好的特征表示方法不会受到多维数据损坏和干扰的影响，引入 DAE 作为特征提取模块可以从混杂的输入数据中重建原始的特征表示，选取最佳特征. DAE 的原理是采用部分破坏原始输入特征的方法，通过从具有噪声和波动的数据输入中重建有效的输入来强制自动编码器的隐藏层捕获输入数据的最佳特征，得到更高质量和鲁棒性的特征表示. 消除噪声不是 DHM 模型的目标，而是通过引入损坏的过程实现特征重构，使特征提取更接近输入数据空间中存在的稳定结构和不变特征，学习更有效的特征表示[8].

无监督学习过程有助于调整深度学习方法中的神经网络参数，对多维特征空间的分布进行优化. 基于上述研究，提出将输入数据空间的特征提取神经网(DAE)和输出空间的无监督学习算法(SOM)在结构上进行组合为降噪分层映射模型，如图 1 所示.

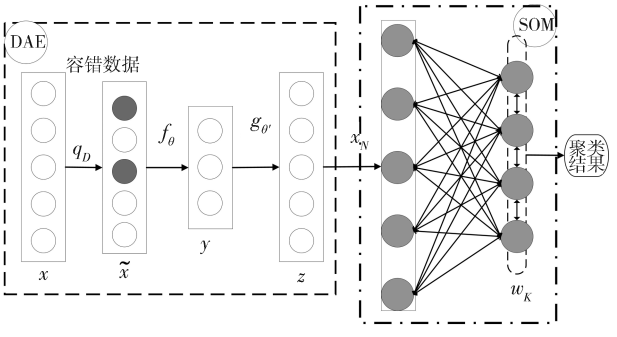


图 1　DHM 算法模型
Fig. 1　Structure of DHM

在图 1 中，第一阶段，容错数据通过特征选择模型的训练，基于 DAE 的特征选择模型通过在训练神经网络的过程中引入随机破坏行为，获取输入数据的分布特征和相关性特征，消除容错数据中破坏数据的噪声使特征选择更接近输入数据空间中的稳定结构和不变特征；第二阶段，无监督聚类模型接受特征选择神经网络的隐藏层输出，捕获并投影从特征选择中提取的特征之间的关系，并将它们同时映射到二维平面上，达到基于拓扑的聚类分析[9].

通过有效地结合降噪自动编码器和自组织映射神经网络，可以在神经网络图中显示输入特征的

非线性组合的影响，并基于最有效特征来实现无监督聚类方案[3].

2.2　特征选择模型

特征选择模型 DAE 可用于从采样子集中选取有效的关联特征重构数据集，如图 1 所示[10]. 输入的容错数据定义如下.

$$\tilde{x}=q_D(x) \tag{1}$$

其中，q_D是破坏函数，在数据采样方案中q_D是使$\tilde{x}$成为x的一个子集的掩码函数.

容错数据向量$\tilde{x}$被映射为隐藏层表示y后经解码输出重构数据z，定义如下.

$$y=f_\theta(\tilde{x}),\ z=g_{\theta'}(y) \tag{2}$$

其中，$\theta=\{W,b_f\}$和$\theta'=\{V,b_g\}$是最佳参数集，$W_{[k\times n]}$为输入权重矩阵，b_f为输入偏置向量，$V_{[n\times k]}$和b_g分别为输出权重矩阵和偏置向量. 由于 DAE 的目的是从容错数据$\tilde{x}$中重构原始数据x，损失函数被定义为原始数据x与重构数据z间的平方误差.

$$j_{\theta,\theta'}=\frac{1}{m}\sum_{i=1}^{m}\|z^{(i)}-x^{(i)}\|_2^2=\frac{1}{m}\sum_{i=1}^{m}\|g_{\theta'}(f_\theta(\tilde{x}^{(i)}))-x^{(i)}\|_2^2 \tag{3}$$

其中，$z^{(i)}$为选取的容错数据$\tilde{x}^{(i)}$经隐藏层映射到输出层的重构数据，基于小批量的梯度下降[11] GD (Gradient Descent)算法用于解决问题并学习参数.

2.3　无监督聚类模型

聚类模型采用自组织特征映射(SOM)构成两层前馈神经网络如图 1 第二部分所示[12]. 左为输入层，由输入结点构成，若输入向量有N个维度则输入端共有N个神经元. 右为竞争层，由K个输出神经元构成，形成一个等间距的二维神经元矩阵，每个神经元对应一个最佳匹配获胜神经元，代表了一类具有相似特征的数据点集群. 所有输入神经元到输出神经元都有权值连接，竞争层的每个神经元同它周围的其他神经元侧向连接.

SOM 的输入数据构成一个包含N维欧几里得向量x的序列$\{x(t)\}$，其中一个整数t表示序列中的一个步骤. $\{m_i(t)\}$为另一个表示连续计算的n维实向量序列的近似模型. i是与m_i相关的网格节点的空间索引，原始 SOM 算法假定以下过程收敛并生成模型所需的有序值.

$$m_i(t+1)=m_i(t)+h_{ci}(t)[X(t)-m_i(t)] \tag{4}$$

其中，$h_{ci}(t)$是邻域函数，下标c是网络中特定节点(获胜者)的索引，即模型$m_i(t)$与$x(t)$具有最小欧几里得距离.

$$c=\arg\min_i\{\|X(t)-m_i(t)\|\} \tag{5}$$

输入数据根据式(5)选择网络中的最佳匹配节点，根据式(4)和式(5)修改了获胜节点以及网络中其空间邻域处的模型. 更新后的模型将更好地与输入匹配. 在不同节点上的修改率取决于领域函数$h_{ci}(t)$的数学形式.

$$h_{ci}(t)=\alpha(t)\exp[-sqdist(c,i)/2\delta^2(t)] \tag{6}$$

其中，$\alpha(t)$是t的单调递减的标量函数. (c,i)是网络中节点c和i之间几何距离的平方，$\delta(t)$为另一个t的单调递减函数. 为了获得足够的统计准确性，每个模型都必须更新. 特别地，δ不能为零，否则将失去其排序能力.

每一个输入数据经过神经网络训练后都要对获胜神经元及其周围节点的权值进行调整以更加匹配输入权值. 稳定后的网络得到与输入数据对应的特征映射空间，实现自动聚类.

3　降噪分层映射(DHM)算法

DHM 模型在统一的体系结构中结合了输入数据空间的特征选择神经网路模型和输出空间自组织特征映射的聚类模型，如图 2 所示.

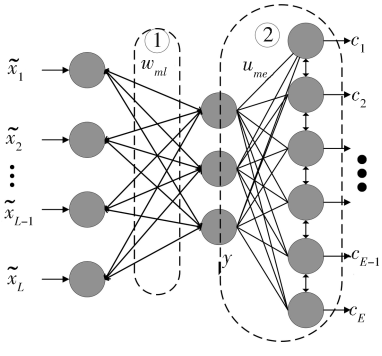


图 2　DHM 算法网络结构

Fig. 2　DHM algorithm network structure

在图 2 中，算法的学习过程分为两个阶段，第一阶段，构建 DAE 神经网络训练得到有效的输入特征空间[13]；在第二阶段，DAE 的隐藏层编码应用于 SOM 的输入，SOM 使用这些分布式特征进行参数调整，实现无监督聚类目标.

3.1　DHM 算法学习

3.1.1　第一阶段学习　如图 2 第一部分所示，输入层和输出层具有相同数量的神经元. 每层神经元的数量与输入样本x的维数L一致，隐藏层的神经元数量为y. 训练开始，输入数据在输入层产生

维数为 L 的容错样本 $\tilde{x}$.

非线性特征在下一层产生利用权重w_{ml}构造的 M 个神经元,表示输入特征$\tilde{x}_l$到隐藏神经元 m 的贡献. 首先,进行仿射变换$z_m=\sum_{l=1}^{L} w_{ml}\tilde{x}_l+b_m$,$b_m$是偏项. 仿射变换$z_m$采用一个被压扁的 s 型正切函数或双曲正切函数 f,得到每个隐藏神经元的激活函数$y_m=f(z_m)$.

在输出层,样本的重构特征由样本特征$\hat{x}_h$通过一组权重$\hat{w}_{hm}$调整的激活函数y_m生成. 这些权重和一组伴生的偏置参数$\hat{b}_h$组合得到仿射映射$\hat{z}_h=\sum_{m=1}^{M}\hat{w}_{hm}y_m+\hat{b}_h$,采用一个压扁函数$\hat{x}_h=f(\hat{z}_h)$生成重构样本$\hat{X}$.

神经网络的参数按照一系列矩阵排列,由$W_{M\times L}$,$b_{M\times 1}$,$\hat{W}_{L\times M}$和$\hat{b}_{L\times 1}$得到紧凑方程表示为

$$Z=W\tilde{X}+b \tag{7}$$

$$y=f(Z)=f(W\tilde{X}+b) \tag{8}$$

$$\hat{Z}=\hat{W}y+\hat{b} \tag{9}$$

$$\hat{X}=f(\hat{Z})=f(\hat{W}y+\hat{b}) \tag{10}$$

其中,f 的定义扩展到适用于各个元素向量,通过设置约束条件$\hat{W}=W^T$($\hat{w}_{hm}=w_{hm}$)得到 W 和$\hat{W}$之间的权重. 加上这个对称假设减少了模型中自由参数的数量. 对于训练样本 X,要求关联的重构误差最小,定义为

$$J(W,b,\hat{b})=\frac{1}{2}\|X-\hat{X}\|^2 \tag{11}$$

这是用于实值输入的平方误差函数,对于二进制或 $X\in[0,1]^L$ 情况,输入交叉熵损失函数代替. 更新网络参数采用梯度下降方法,以最大程度地减少重构误差,步骤如下.

$$W^{(\text{next})}=W-\alpha\nabla_W J(W,b,\hat{b}) \tag{12}$$

$$b^{(\text{next})}=b-\alpha\nabla_b J(W,b,\hat{b}) \tag{13}$$

$$\hat{b}^{(\text{next})}=\hat{b}-\alpha\nabla_{\hat{b}} J(W,b,\hat{b}) \tag{14}$$

其中,α 为学习率;W,b 和$\hat{b}$是偏导数梯度下降方程.

3.1.2　第二阶段学习　如图 2 所示,考虑到拓扑上的低维有序神经元,E 为 SOM 神经网络的竞争层神经元总数. 每个神经元 e 与 SOM 的可调参数码本矢量u_e相互关联,可以看作点u_{me}属于每个单独隐藏层y_m的分布表示,码本矢量可以排列成矩阵$U_{M\times E}$.

在第二阶段,训练算法运用梯度下降在建模任务中寻求最小值,将成本函数[14]定义为

$$K(W,b,U)=\min_d\sum_{e=1}^{E}h_{de}\frac{1}{2}\|y-U_e\|^2 \tag{15}$$

其中,h_{de}是邻域函数,它的宽度和形状控制表面的弹性神经元. K 引入一个加权平均过程表示每个码本矢量和给定隐藏层间的距离. 权重系数的作用是加强一些规则的限制,成本函数可以相应表示为

$$K(W,b,U)=\sum_{c=1}^{E}N(y,c)\sum_{e=1}^{E}h_{ce}\frac{1}{2}\|y-U_e\|^2 \tag{16}$$

$$N(y,c)=\begin{cases}1, & c=\operatorname{argmin}_d\sum_{e=1}^{E}h_{de}\|y-U_e\|^2\\ 0, & \text{otherwise}\end{cases} \tag{17}$$

其中,h_{ce}为神经元在晶格上的几何坐标,对于有限的输入样本集,成本函数是连续的,用于最小化成本函数的梯度下降步骤定义为

$$W^{(\text{next})}=W-\eta\nabla_W K(W,b,U) \tag{18}$$

$$b^{(\text{next})}=b-\eta\nabla_b K(W,b,U) \tag{19}$$

$$U^{(\text{next})}=U-\theta\nabla_U K(W,b,U) \tag{20}$$

其中,θ 和 η 是标量学习率因子,学习率 η 调整权重,偏差应该传递比学习率 θ 低的值来调整映射. 因此,在第二阶段开始时,η 和 θ 必须至少相差一个数量级以缓和训练失衡.

3.2　DHM 算法实现

算法在第一阶段学习产生非线性空间,在第二阶段期间使用不同的优化目标进一步完善特征选择. 因为无法评估用于无监督学习的表示空间学习阶段之间的比例平衡无法精确调整. 为了解决两阶段交错训练的问题,首先由式(12)~(14)调整 DEA 的参数并表示特征空间. 再根据上一阶段固定和不变的表示空间式(20)来调整 SOM,然后进行所有模型参数的组合并行微调. 最后由式(18)~(20)对模型进行训练,从而生成一个更加有效的特征选择模型,避免了表示空间的广泛失真,有利于神经映射.

更新 DHM 参数的过程的每个特定步骤如下,第一阶段的迭代参数表示为

$$w_{ij}^{(\text{next})}=w_{ij}+a\left[(x_j-\hat{x}_j)f'(\hat{z}_j)y_i+\sum_{h=1}^{L}(x_h-\hat{x}_h)f'(\hat{z}_h)w_{ih}f'(z_i)\tilde{x}_j\right],$$
$$1\leqslant i\leqslant M,1\leqslant j\leqslant L \tag{21}$$

$$b_i^{(\text{next})}=b_i+a\sum_{h=1}^{L}(x_h-\hat{x}_h)f'(\hat{z}_h)w_{ih}f'(z_i),$$

$$1 \leqslant i \leqslant M \tag{22}$$

$$\hat{b}_j^{(\text{next})} = \hat{b}_j + a(x_j - \hat{x}_j) f'(\hat{z}_j), 1 \leqslant j \leqslant L \tag{23}$$

其中，权重参数w_{ij}的第二项包含对应的激活函数y_i，我们通过使用链式规则得到$\hat{x}_h = f(\hat{z}_h)$和$\hat{z}_h = \sum_{m=1}^{M} w_{mh} y_m + \hat{b}_h$用于计算隐藏层偏置参数$b_j$和输出层偏置参数$\hat{b}_j$，$f'$为 f 的总倒数. 第二阶段的遍历参数和必需的重复次数定义为

$$c = \underset{d}{\operatorname{argmin}} \sum_{e=1}^{E} h_{de} \sum_{m=1}^{M} (y_m - u_{me})^2 \tag{24}$$

$$w_{ij}^{(\text{next})} = w_{ij} + \eta \sum_{e=1}^{E} h_{ce}(u_{ie} - y_i) f'(z_i) x_j,$$
$$1 \leqslant i \leqslant M, 1 \leqslant j \leqslant L \tag{25}$$

$$b_i^{(\text{next})} = b_i + \eta \sum_{e=1}^{E} h_{ce}(u_{ie} - y_i) f'(z_i),$$
$$1 \leqslant i \leqslant M \tag{26}$$

$$u_{ie}^{(\text{next})} = u_{ie} + \theta h_{ce}(y_i - u_{ie}),$$
$$1 \leqslant i \leqslant M, 1 \leqslant e \leqslant E \tag{27}$$

由链式规则有$y_m = f(z_m)$和$z_m = \sum_{l=1}^{L} w_{ml} \tilde{x}_l + b_m$，$u_{ie}$为码本向量参数. 在学习过程之后，结果如图 2 中描绘的 DHM 所示，所有参数都已更新和调整. 输入样本在 DHM 平面上的图像定义为在隐藏层表示与相应的码本矢量之间产生最小加权平方欧式距离的神经元[12]. 加权是指在 DHM 网格的拓扑上定义的邻域内核，最佳匹配的获胜者神经元由下式给出.

$$y_m = f\left(\sum_{l=1}^{L} w_{ml} x_l + b_m\right), 1 \leqslant m \leqslant M \tag{28}$$

$$c = \underset{d}{\operatorname{argmin}} \sum_{e=1}^{E} h_{de} \sum_{m=1}^{M} (y_m - u_{me})^2 \tag{29}$$

最后，如果单峰邻域函数的半径足够窄，几乎可以涵盖最接近的领域，那么先前检测到的最佳匹配神经元将与 SOM 的获胜者神经元重合.

$$c = \underset{d}{\operatorname{argmin}} \| y - U_d \|^2 \tag{30}$$

只要适用邻域大小条件两个表达式可以互换使用. 基于以上分析，迭代算法解决无监督学习问题的伪代码如算法 1 所示.

算法 1：降噪分层映射算法(DHM)

输入：容错数据，DHM 网络参数{神经网络权重w_{ml}，$\hat{w}_{lm}$，学习率 α，码本矢量u_{me}，标量学习率因子 θ 和 η}

输出：激活函数$y_m = f(z_m)$，聚类模型参数，最佳匹配的获胜神经元 c.

1) 初始化：学习率 α，权重w_{ml}，w_{lm}，码本矢量u_{me}，获胜神经元c_E.
2) Repeat
3) for $i = 1, 2, \cdots, i_{\max}$ do
4) 收到$\tilde{x}_i$，
5) 通过$\tilde{x}_i$由式(12)～(14)学习训练特征选择模型$\{W, b, \hat{b}\}$.
6) 根据特征空间由式(20)调整 SOM，
7) 通过式(18)～(20)最小化成本函数 $K(W, b, U)$并更新学习率因子.
8) 由式(28)和式(29)得到容错数据$\tilde{x}_i$对应的最佳匹配获胜神经元 c_E.
9) end for

在此方法的每个步骤中，所有变量均根据其相应公式进行更新，此过程迭代进行，直到变量收敛为止.

3.3　DHM 算法分析及评价

为了验证所提出的算法，内部和外部评估标准共同构成了算法评价体系[15]. 外部聚类标准包含标准化互信息，聚类纯度. 标准化互信息要求定义一些额外的数量 a，s_p代表集群类别中第 p 组的样本数，s_t代表现有分类第 t 类的样本数. $s_{p,t}$表示同时出现在第 p 分区和第 t 类别的样本数.

$$NMI = \frac{\sum_{p=1}^{P} \sum_{t=1}^{T} s_{p,t} \log\left(S \frac{s_{p,t}}{s_p s_t}\right)}{\sqrt{\left(\sum_{p=1}^{P} s_p \log \frac{s_p}{S}\right)\left(\sum_{t=1}^{T} s_t \log \frac{s_t}{S}\right)}} \tag{31}$$

NMI 的值域为[0，1]，较低的值表示较高的不确定性. 纯度是聚类分析中广泛使用的外部度量.

$$PUR = \frac{1}{S} \sum_{p=1}^{P} \max_{1 \leqslant t \leqslant T} |s_p \cap s_t| \tag{32}$$

聚类性能的内部度量包含均方量化误差和 DB 度量[14]. 量化误差通过以下各项来确定网络对输入和表示的适应性.

$$QE = \frac{1}{S} \sum_{s=1}^{S} \| y^{(s)} - U_c^{(s)} \| \tag{33}$$

DB 度量表示最接近簇之间的成对相似度平均值的估计值.

$$DB = \frac{1}{p} \sum_{i=1}^{P} \max_{1 \leqslant j \leqslant P, i \neq j} \frac{CD_i + CD_j}{MD_{ij}} \tag{34}$$

其中，CD_i是集群 i 的离散度. MD_{ij}是集群 i 和 j 的码本向量值在[0，+∞]范围内，较高的离散度和相似群集之间的距离较大得到较高的 DB 值，反应了

群集和神经映射的欠佳.

$$CD_i = \left\{ \frac{1}{S_i} \sum_{j=1}^{S_i} \| y^{(i)} - U_i \|^q \right\}^{1/q} \tag{35}$$

$$MD_{ij} = \left\{ \sum_{k=1}^{M} | u_{ki} - u_{kj} |^r \right\}^{1/r} \tag{36}$$

其中,S_i是集群 i 的样本数;u_i是相同集群的质心. 当 $r=2$且 $q=1$ 时,MD_{ij}是质心和CD_i之间的欧氏距离.

4　仿真分析

4.1　仿真环境和方法

为了验证算法的多维聚类性能,选用 UCI Machine Learning Repositor 中多维数据集手写数字的光学识别 Optdigits[16]评估 DHM 算法解决实际问题的能力. 表 1 为仿真数据集信息,仿真环境为:python 3.7;Win10 64 位操作系统;2.6 GHz CPU;8 G 内存.

表 1　数据集信息

Tab. 1　Data set information

数据集	实例数	属性数	类别数
Optdigits	5620	64	10

为了验证所提出的算法,实验基于 Optdigits 数据集,交替固定隐藏层神经元数量和噪声类型与参数,观察相对变量变化下的聚类性能.

4.2　外部聚类性能影响分析

为了研究算法外部聚类性能,在 Optidigits 数据集中评估 DHM 算法与 AESOM,DCSOM 和 S-SOM 算法在固定噪声参数研究隐藏层神经元数量变化对外部聚类性能的影响,以及固定隐藏层神经元数量研究不同噪声参数类型变化下的外部聚类评价指标结果. 仿真均取 20 次重复实验的平均值,结果如图 3 和图 4 所示.

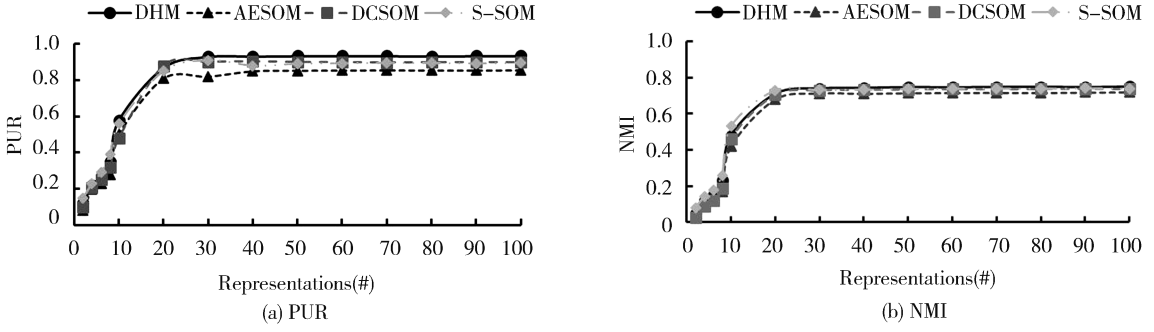


图 3　算法外部聚类指标对比,高斯噪声的标准偏差保持恒定且等于 0.5

Fig. 3　Algorithm external clustering index comparison, with Gaussian noise's standard deviation kept constant and equal to 0.5

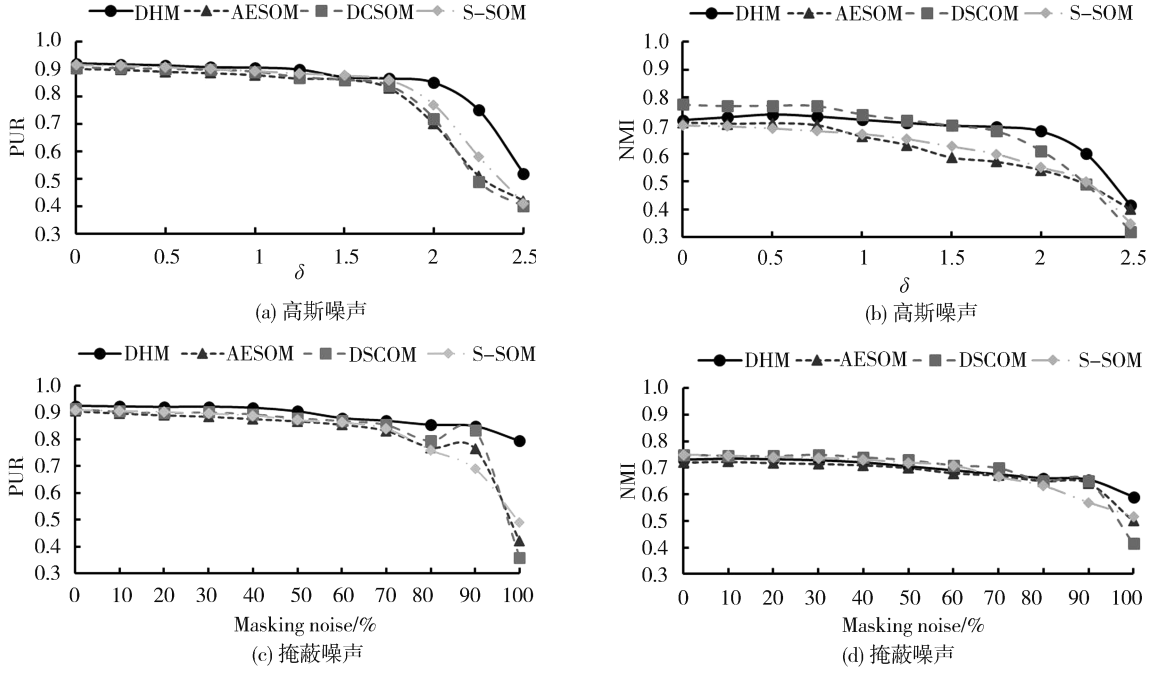


图 4　具有 32 个隐藏层神经元的算法外部聚类指标对比

Fig. 4　Comparison of external clustering indicators of algorithms with 32 hidden neurons

从图 3 可以观察到，固定高斯噪声的标准偏差为 0.5 的情况下，4 种算法的整体趋势随着隐藏层神经元数量的提高拥有更高的聚类精度. 同对比算法相比 DHM 算法在外部聚类评价指标 PUR 和 NMI 上性能差异较小，但当隐藏层表示为 10＃～20＃时，DMH 和 S-SOM 算法的外部聚类性能提升同 AESOM 和 DCSOM 算法相比较为明显，但随着隐藏层神经元数量的增加，DHM 算法具有更高的聚类精度.

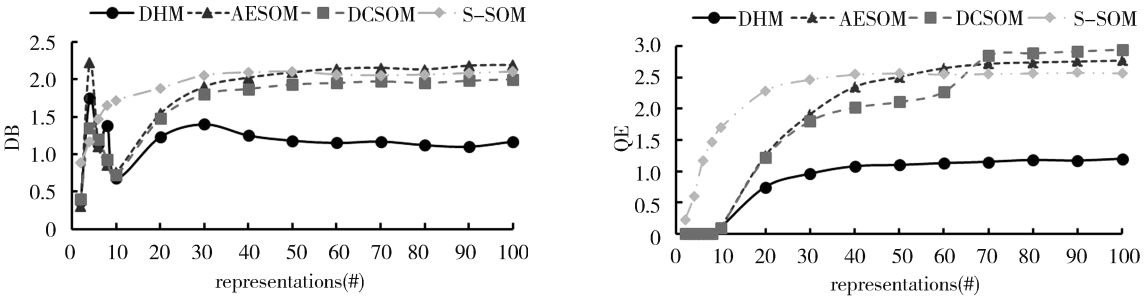
从图 4a 和 4b 可以观察到，当神经元数量固定为 32 的情况下，随着高斯噪声标准偏差的增大，4 种算法的外部聚类性能.

当噪声失真在 0～1.5δ 时，4 种算法的 PUR 和 NMI 性能指标没有显著差异. 但当高斯噪声在 1.5δ～2.5δ 时，同对比算法相比 DHM 算法在噪声外部聚类评价指标 PUR 和 NMI 上性能更为优异，且随着噪声失真的变大，DHM 算法能够应对更严重的噪声失真提供更高的聚类精度. 另一方面，从图 4c 和 4d 可以观察到，当掩蔽噪声的破坏百分比在 0%～80%情况下，4 种算法的 PUR 和 NMI 指标性能差异较小，显示了 4 种算法均有较强的抗噪性能. 但当掩蔽噪声破坏百分比达到 90%～100%时，DHM 算法同对比算法相比显示出了更高的聚类精度.

交替固定参数仿真结果表明，DHM 算法在 SOM 模型中引入降噪组件实现特征重构，插入经过调参的隐藏层神经元对聚类分析提供了更优异的性能，同对比算法相比在交替固定隐藏层表示和噪声参数情况下均能表现出更高的聚类精度和抗噪鲁棒性.

4.3　内部聚类性能影响分析

为了研究算法的内部聚类性能，在 Optidigits 数据集中评估 DHM 算法与 AESOM，DCSOM 和 S-SOM 算法的内部聚类评价指标结果. 仿真均取 20 次重复实验的平均值，结果如图 5 和图 6 所示.



(a) DB　　(b) QE

图 5　算法内部聚类指标对比，高斯噪声的标准偏差保持恒定且等于 0.5

Fig. 5　Algorithm internal clustering index comparison, with Gaussian noise's standard deviation kept constant and equal to 0.5

从图 5 可以观察到，固定高斯噪声的标准偏差为 0.5 的情况下，相较对比算法 DHM 算法在内部聚类评价指标 DB 和 QE 上性能更为优异. 随着隐藏层表示的增大，四种算法的内部聚类指标总体趋势为在 0＃～40＃的隐藏层表示阶段变化上升，在 40＃～100＃阶段逐渐达到平稳，S-SOM 算法的 DB 和 QE 性能均较快达到平稳，分析是由于权重对指数距离的影响使得集群分离性和紧凑型较差更小的 DB 和 QE 值显示出 DHM 算法更好的集群分离性和紧凑性.

从图 6a 和 6b 可以观察到，当神经元数量为 32 个的情况下，在高斯噪声失真的整个变化过程中，不论是 DB 指标还是 QE 指标，DHM 同对比算法相比均表现出更小值，显示出算法更好的集群分离性. 从图 6c 和 6d 可以观察到，当神经元数量为 32 个的情况下，随着掩蔽噪声破坏百分比的增加，在 0%～80%阶段 4 种算法的 DB 和 QE 指标均呈现缓慢的平稳降低状态，在 80%～100%下出现波动变化后快速降低的情况. DHM 算法在掩蔽噪声变化的整个过程中，均有更低的 DB 和 QE 值，DHM 算法明显改善了集群组织的可分离性，具有更好的内部聚类性能.

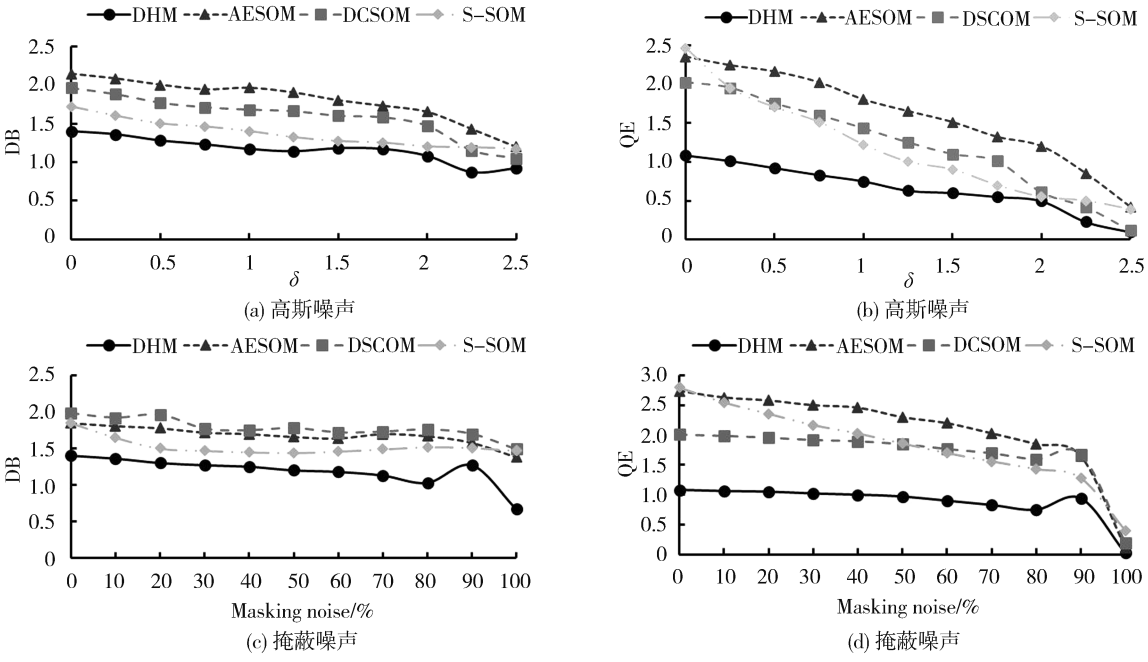


图 6　具有 32 个隐藏层神经元的算法内部聚类指标对比

Fig. 6　Comparison of internal clustering indicators of algorithms with 32 hidden neurons

结合两次仿真实验,DHM 算法的输入特征空间的隐藏层表示更加紧凑,保留了原始数据关联特征实现良好的集群可分离性,具有更高的聚类准确性和神经网络拟合性. 在基准数据集线圈和耶鲁人脸数据库的仿真实验也同样验证了 DHM 算法的聚类准确性和有效性[17].

5　结　论

采用深度神经网络学习训练,可以选取有效的特征属性表征数据间的关联信息进行无监督聚类分析,本文研究提出一种降噪分层映射算法(DHM). 首先,基于降噪自动编码器的神经网络从原始数据中学习训练得到激活函数特征空间实现特征重构. 其次,特征空间用于训练自组织特征映射神经网,通过计算最小加权平方欧式距离来寻找最佳匹配获胜神经元. 最后,结合已调组件进行共同训练,优化权重系数和偏置参数,实现有效的多维聚类分析. 仿真结果同对比算法相比,DHM 算法在聚类精确性和有效性方面均有提升. 下一步,面对更为复杂的数据结构和分析需求,将研究更有效深度学习方法.

参考文献:

[1]　刘凯, 方勇, 张磊, 等. 基于图卷积网络的恶意代码聚类 [J]. 四川大学学报: 自然科学版, 2019, 56: 654.

[2]　Mousavi S M, Zhu W, Ellsworth W, *et al*. Unsupervised clustering of seismic signals using deep convolutional autoencoders[J]. IEEE Geosci Remote S, 2019, 16: 1693.

[3]　Guo J Q, Liu Y Z X, Zhang L F, *et al*. Driving behavior style study with a hybrid deep learning framework based on GPS data[J]. Sustainability, 2018, 10: 2351.

[4]　Aly S, Almotairi S. Deep convolutional self-organizing map network for robust handwritten digit recognition [J]. IEEE Access, 2020, 8: 107035.

[5]　Pd A, Ldg B, Rm A. Smoothed self-organizing map for robust clustering[J]. Inform Sciences, 2020, 512: 381.

[6]　Li T, Dong H. Unsupervised feature selection and clustering optimization based on improved differential evolution[J]. IEEE Access, 2019, 7: 140438.

[7]　邓描, 刘强, 陈洪刚, 等. 一种基于特征正则约束的异常检测方法 [J]. 四川大学学报: 自然科学版, 2020, 57: 1077.

[8]　Pulgar F J, Charte F, Rivera A J, *et al*. ClEnDAE: a classifier based on ensembles with built-in dimensionality reduction through denoising autoencoders [J]. Inform Sciences, 2021, 565: 146.

[9]　Bassani H F, Araujo A F R. Dimension selective self-organizing maps with time-varying structure for subspace and projected clustering [J]. IEEE T Neur Net Lear, 2015, 26: 458.

[10]　Yu T, Wang X, Shami A. UAV-Enabled spatial

data sampling in large-scale iot systems using denoising autoencoder neural network [J]. IEEE Internet Things J, 2019, 6: 1856.

[11] Hao W. A gradient descent method for solving a syste of nonlinear equations[J]. Appl Math Lett, 2021, 112: 106739.

[12] Kohonen T. Essentials of the self-organizing map [J]. Neur Net, 2013, 37: 52.

[13] Vincent P, Larochelle H, Lajoie I, *et al*. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. J Mach Learn Res,2010, 11: 3371.

[14] Gepperth A. An energy-based SOM model not requiring periodic boundary conditions [C]//Proceedings of the 2017 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM). Nancy, France: IEEE, 2017.

[15] Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions [J]. J Mach Learn Res, 2002, 3: 583.

[16] Bache K, Lichman M. UCI machine learning repository [DB/OL]. [2020-11-08]. http://archive. ics. uci. edu/ml.

[17] Nene S A, Mayar S K, Murase H. Columbia object image library (COIL-20): CUCS-005-96[R]. New York, USA: COLL, 1996.

引用本文格式：

中　文：刘云，张轶，郑文凤.降噪分层映射算法在多维聚类分析中的优化研究[J]. 四川大学学报：自然科学版，2022，59：013001.

英　文：Liu Y, Zhang Y, Zheng W F. Optimization research of denoised hierarchical mapping analysis for multidimensional cluster analysis [J]. J Sichuan Univ: Nat Sci Ed, 2022, 59: 013001.