一种基于ELM-AE特征表示的谱聚类算法

王丽娟， 丁世飞

引用本文:
王丽娟, 丁世飞. 一种基于ELM-AE特征表示的谱聚类算法[J]. 智能系统学报, 2021, 16(3): 560-566.
WANG Lijuan, DING Shifei. A spectral clustering algorithm based on ELM-AE feature representation[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(3): 560-566.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202005021

DOI: 10.11992/tis.202005021

一种基于 ELM-AE 特征表示的谱聚类算法

王丽娟[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 信息工程学院, 江苏 徐州 221114)

摘 要: 在实际应用中, 数据点中包含的冗余特征和异常值 (噪声) 严重影响了聚类中更显著的特征的发现, 大大降低了聚类性能。本文提出了一种基于 ELM-AE (extreme learning machine as autoencoder) 特征表示的谱聚类算法 (spectral clustering via extreme learning machine as autoencoder, SC-ELM-AE)。ELM-AE 通过奇异值分解学习源数据主要特征表示, 使用输出权值实现从特征空间到原输入数据的重构; 再将该特征表示空间作为输入进行谱聚类。实验表明, 在 5 个 UCI 数据集验证中, SC-ELM-AE 算法性能优于传统的 K-Means、谱聚类等现有算法, 特别是在复杂高维数据集 PEMS-SF 和 TDT2_10 上, 聚类平均精确度均提高 30% 以上。

关键词: 谱聚类; 特征表示; 极限学习机; 自编码器; 极限学习机自编码器; 机器学习; 聚类分析; 数据挖掘
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2021)03−0560−07

中文引用格式: 王丽娟, 丁世飞. 一种基于 ELM-AE 特征表示的谱聚类算法 [J]. 智能系统学报, 2021, 16(3): 560–566.
英文引用格式: WANG Lijuan, DING Shifei. A spectral clustering algorithm based on ELM-AE feature representation[J]. CAAI transactions on intelligent systems, 2021, 16(3): 560–566.

A spectral clustering algorithm based on ELM-AE feature representation

WANG Lijuan[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 Engineering, Xuzhou College of Industrial Technology, Xuzhou 221114, China)

Abstract: In practice, redundant features and outliers (noise) in data points heavily influence the discovery of more prominent features in clustering and significantly impair clustering performance. In this study, we propose a spectral clustering (SC) based on extreme machine learning as autoencoder (ELM-AE) feature representation (SC-ELM-AE). ELM-AE learns the principal feature representation of the source data via singular value decomposition and uses the output weights to realize reconstruction from feature representation space to the original input data. The reconstructed feature representation space is fed to the SC as input. The experimental results show that the proposed algorithm is 30% more accurate in the average clustering than the conventional K-means, SC, and other existing algorithms in the verification of five UCI datasets, particularly on complex high-dimensional datasets, such as PEMS-SF and TDT2_10.

Keywords: spectral clustering; feature representation; extreme machine learning; auto-encoder; extreme learning machine as autoencoder; machine learning; clustering analysis; data mining

聚类[1-3] 是一种将数据集划分为若干组或类, 使类内相似性最大, 类间相似性最小的方法。现有文献提出, 传统的聚类方法有 k 均值算法 (k-means)[4]、FCM 算法[5]、子空间聚类等。它们虽然简单, 但缺乏处理复杂数据结构的能力, 当样本为凸型时, 这些算法对数据的处理有较好的效果, 当样本空间为非凸时, 算法容易陷入局部最优。为解决这一难题, 谱聚类应运而生, 它不受样本空间形状限制, 聚类结果为全局最优解。在这些算法中, 由于谱类算法便于实现、性能良好, 已被广泛应用于图像分割、信息检索、人脸识别等领域。

收稿日期: 2020−05−17.
基金项目: 国家自然科学基金项目 (61672522, 61976216); 江苏省高校哲学社会科学研究项目 (2019SJA1013); 江苏高校 “青蓝工程”.
通信作者: 丁世飞. E-mail: dingsf@cumt.edu.cn.

谱聚类来源于图论中的最小切割问题[6-8]。众所周知, 数据通过非线性变换, 将数据映射到高维特征空间后, 数据将变得线性可分。因此, 可以通过各种非线性变换, 例如基于核的方法, 提高高维特征空间中输入数据的特征表示能力。然而, 以往的谱聚类方法一般只注重数据降维、优化时间复杂度等, 而不是进一步提高数据特征的表示能力。在聚类任务中, 数据的特征表示是至关重要的, 通过反复的实验和一系列的工作, 获得良好的数据特征表示是提高聚类准确度的保证。

ELM[9-11] 最初用于训练"广义"单隐层前馈神经网络 (SLFN), 具有快速学习、良好的泛化能力和特征表示能力, 被广泛应用于各种机器学习任务, 如回归和分类等[12-14]。近年来, ELM 已经扩展到聚类, 一般是在 ELM 获得的嵌入特征空间中进行聚类。ELM 的隐层参数选择是随机的, 和输入数据无关, 与常用的反向传播训练算法相比, ELM 最大限度地减小了训练误差和输出权的范数。根据 Bartlett 理论[15], 最小化输出权值范数将产生更好的泛化能力。

自编码器 (Auto Encoder AE)[16-17] 是深度学习中的一种无监督学习方法, 能够从大量数据中自动学习到数据中的有效特征[18-20]。传统自编码器就是一种在输出层重建输入数据并且结构对称的神经网络。常见的自编码器有降噪自编码器、稀疏自编码器、收缩自编码器和卷积自编码器等[21]。使用自编码器进行特征提取要比特征分解快, 并让目标值等于输入值, 是一种尽可能复现输入信号的神经网络。

基于以上启发, 本文将极限学习机 (ELM) 和自编码器 (AE) 结合起来改进谱聚类, 提出了一种基于 ELM-AE 嵌入特征空间的谱聚类算法。

1　相关工作

1.1　谱聚类算法

谱聚类的概念起源于谱划分, 将数据聚类转化为无向图的多路划分问题求解, 尤其适用于数据集非凸的情况[22-23]。假设数据集有 n 个数据点, 目标是将所有数据点划分到 c 个簇中。谱聚类[24] 首先将数据点看作图的顶点, 根据数据点成对相似性, 先用欧氏距离计算距离矩阵, 再使用高斯核函数将距离矩阵构造为相似矩阵并计算出所对应的拉普拉斯矩阵, 谱方法是基于相似拉普拉斯矩阵的特征值和特征向量进行聚类的方法, 将这些特征向量构造为低维的特征子空间, 再在这个特征子空间上使用诸如 k-means 的聚类方法进行聚类, NJW 谱聚类算法[24] 是一个典型的谱聚类算法。下面简单介绍该方法的原理。

给定一组数据集 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$, 要把这些数据点分成 k 类, 使用谱聚类的一般过程是:

1) 根据一定的相似矩阵生成方式构建样本的相似矩阵, $\boldsymbol{W}$ 中的每一项表示每一对点之间的相似性:

$$\boldsymbol{W}_{ij}=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2)$$

式中: 当 $i=j$ 时, $\boldsymbol{W}_{ij}=0$。

2) 根据相似矩阵 $\boldsymbol{W}$ 计算度矩阵 $\boldsymbol{D}$, $\boldsymbol{D}$ 是一个对角矩阵, 它的非零元素是 $\boldsymbol{W}$ 的所有行元素 (或者列元素) 的总和:

$$\boldsymbol{d}_{ii}=\sum_j \boldsymbol{w}_{ij}$$

3) 根据 $\boldsymbol{D}$ 和 $\boldsymbol{W}$ 计算出拉普拉斯矩阵 $\boldsymbol{L}$, 定义拉普拉斯矩阵有很多种方法。非标准化的拉普拉斯矩阵为

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$$

4) 对 $\boldsymbol{L}$ 进行标准化会产生更好的聚类效果, 因此一般将拉普拉斯矩阵标准化形式表示为

$$\boldsymbol{L}_{\text{norm}}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$$

还可以表示为

$$\boldsymbol{L}_{\text{norm}}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$$

5) 接下来计算标准化后的拉普拉斯矩阵的前 k 个最小特征值对应的特征向量 $\boldsymbol{f}$, 将各自对应的特征向量 $\boldsymbol{f}$ 组成的矩阵按行标准化, 最终组成 $n\times k$ 维的特征矩阵 $\boldsymbol{F}$;

6) 将 $\boldsymbol{F}$ 中的每一行作为一个 k 维的样本 (共 n 个样本), 用 k-means 算法进行聚类得到最终聚类结果。

1.2　极限学习机自编码器

极限学习机自编码器[19] 是一种既可以再现输入数据也可以自行编码的神经网络方法, 具有极限学习机的计算速度快、精确率高等特点。与传统的极限学习机相似, ELM-AE 包含输入层、单隐含层以及输出层; 不同的是, ELM-AE 的输出层与输入层相等。给定一个训练样本, ELM-AE 的模型结构包括 M 个输入层节点、J 个隐层节点和 M 个输出层节点。ELM-AE 的隐含层输出可以用式 (1) 表示:

$$\boldsymbol{h}=\boldsymbol{g}(\boldsymbol{a}\boldsymbol{x}+\boldsymbol{b}),\ \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}=\boldsymbol{I},\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}=1 \tag{1}$$

式中: $\boldsymbol{a}$ 是输入层和隐含层之间的输入; $\boldsymbol{b}$ 是隐含层的偏置。隐含层输出与 ELM-AE 的输出层之间的数值关系可以用式 (2) 表示:

$$\boldsymbol{h}(\boldsymbol{x}_i)\boldsymbol{\beta}=\boldsymbol{x}_i^{\mathrm{T}},\ i=1,2,\cdots,N \tag{2}$$

式中: $\boldsymbol{\beta}$ 是连接隐含层和输出层的输出权值; $\boldsymbol{g}(\cdot)$ 是激活函数。ELM-AE 的目标是通过最小化正则

化最小二乘估计成本函数得到输出权重，其公式为

$$\min_{\boldsymbol{\beta}} \ \boldsymbol{L}_1 = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{\beta}\|^2 \tag{3}$$

式中：C 是平衡经验风险与结构风险的参数。通过取 $\boldsymbol{L}_1$ 对 $\boldsymbol{\beta}$ 的偏导数并令其等于零，则输出权值矩阵 $\boldsymbol{\beta}$ 为

$$\boldsymbol{\beta} = \begin{cases} \left(\dfrac{\boldsymbol{I}}{C} + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{X}, & N \geqslant J \\ \boldsymbol{H}^{\mathrm{T}}\left(\dfrac{\boldsymbol{I}}{C} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{X}, & N < J \end{cases} \tag{4}$$

2 ELM-AE 特征空间的谱聚类算法

前面从理论层面分析了谱聚类及 ELM-AE 的特征提取方法。传统的谱聚类是将原始数据样本作为聚类初始值，而在实际应用中数据通常冗余且复杂，能够从原始数据中充分挖掘内在信息，通过机器学习算法进行特征学习，使用获得的数据特征空间进行谱聚类将有效提高聚类质量。ELM 可以随机初始化输入权重和偏置并得到相应的输出权重，然而随机生成的输入权值和偏差会导致 ELM 隐层的输出不能很好地代表原始样本的特征。ELM-AE 是一种能够重建输入信号的人工神经网络算法，和自动编码器一样，ELM-AE 无需迭代可以获得原始样本的主要特征，与传统的 ELM 不同的是，ELM-AE 通过选择随机权重和随机偏差正交，可以获得更高的性能。通过 ELM-AE 的输出 $\boldsymbol{\beta}$ 实现从特征空间到原始输入数据的转换，使得输出数据等于输入数据。

本文提出的基于 ELM-AE 嵌入特征空间的谱聚类是将 ELM-AE 的特征空间作为聚类原始值，从而提高聚类性能。SC-ELM-AE 模型包含输入层、单隐层和输出层，其模型结构如图 1，也具有 M 个输入层节点，J 个隐藏层节点和 M 个输出层节点，然后如图所示通过 ELM-AE 获得输出层权值 $\boldsymbol{\beta}$，ELM-AE 输出层的嵌入特征 (embedding feature,EF) 可由式 (5) 计算：

$$\mathbf{EF} = f\left(\boldsymbol{X}\boldsymbol{\beta}^{\mathrm{T}}\right) \tag{5}$$

然后采用归一化谱聚类算法将 ELM-AE 的嵌入特征 (EF) 作为聚类输入数据点进行聚类。SC-ELM-AE 算法流程如图 1 所示。基于 ELM-AE 嵌入特征空间的谱聚类算法 (SC-ELM-AE) 流程详见算法 1。

算法 1 基于 ELM-AE 嵌入特征空间的谱聚类

输入 数据集样本$(x_1, x_2, \cdots, x_n)$，参数 $\boldsymbol{a}$ 和 $\boldsymbol{b}$；

输出 基于 **EF** 空间的 N 个样本的 k 个簇。

1) 初始化由 N 个隐藏层神经元组成的 ELM-AE 网络，计算隐藏层的输出 $\boldsymbol{H}_{\mathrm{ELM-AE}} \in \mathbf{R}^{n\times n}$，激活函数使用 Sigmod(·)；

2) 使用式 (3)、(4) 计算 ELM-AE 的输出层权值；

3) 使用式 (5) 计算嵌入特征 EF，激活函数使用 Sigmod(·)；

4) 得到嵌入特征样本 $\mathbf{EF} = (\mathbf{EF}_1, \mathbf{EF}_2, \cdots, \mathbf{EF}_n)$，初始化高斯相似度函数的参数 σ，k 个聚类；

5) 在 **EF** 样本空间上，使用高斯相似度函数 $\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\left\|f(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) - f(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j)\right\|^2\right)/2\sigma^2\right)$ 构造相似度矩阵 $\boldsymbol{S} \in \mathbf{R}^{n\times n}$，当 $i = j$ 时，$\boldsymbol{S}_{ij} = \boldsymbol{G}(\mathbf{EF}_i, \mathbf{EF}_j)$，否则 $\boldsymbol{S}_{ij} = 0$；

6) 构造归一化对称拉普拉斯矩阵 $\boldsymbol{L}_{\mathrm{norm}} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$，其中 $\boldsymbol{D}$ 是对角矩阵，其 (i, j) 元素是 S 的第 i 行元素的和；

7) 将矩阵 $\boldsymbol{L}_{\mathrm{norm}}$ 前 k 个最大特征向量作为列构造矩阵 $\boldsymbol{Y} \in \mathbf{R}^{n\times k}$；

8) 归一化矩阵 $\boldsymbol{Y}$ 的行，得到矩阵 $\boldsymbol{U} \in \mathbf{R}^{n\times k}$，其中 $\boldsymbol{u}_{ij} = \boldsymbol{y}_{ij}/\left(\sum_k \boldsymbol{y}_{ik}^2\right)^{1/2}$；

9) 利用 k-means 算法对矩阵 $\boldsymbol{U}$ 的行向量聚类，得到 k 个簇。

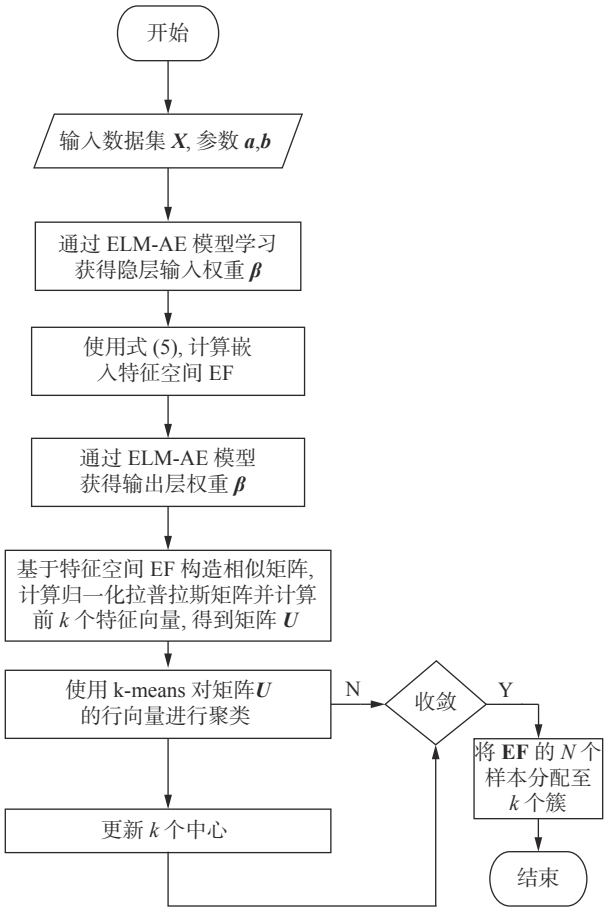


图 1 SC-ELM-AE 算法流程

Fig. 1 Flow of SC-ELM-AE algorithm

3　实验结果与分析

3.1　实验环境及数据集说明

为了验证基于 ELM-AE 嵌入特征空间的谱聚类算法的有效性, 选取了 UCI 机器学习数据库中的 5 个常用数据集进行验证测试, 数据集的基本特征如表 1 所示。

表 1　实验中使用的 UCI 数据集

Table 1　UCI datasets used in the experiments

数据集	样本数	维度	类
WDBC	569	30	2
Ionosphere	351	34	2
Isolet	7797	617	26
PEMS-SF	440	138672	7
TDT2_10	653	36771	10

实验在 MATLAB R2016b 环境下实现, 运行在 Windows 10 上, 实验中使用的计算机 CPU 型号为 Inter(R)Core(TM)i5-8250U @1.60 GHz, 内存为 8 GB。本文所有实验中, ELM-AE 模型包含输入层、输出层、隐藏层, 为了方便实验, 隐藏层、输出层的激活函数都采用 sigmoid 函数。

实验使用的对比算法分别为: 传统谱聚类[25] (spectral clustering, SC)、k-means 聚类、基于无监督极限学习机 (unsupervised extreme learning machine US-ELM) 的 K-means 聚类、基于 ELM-AE 嵌入特征的无监督极限学习机[26](unsupervised extreme learning machine based on embedded features of ELM-AE, US-EF-ELM)k-means 聚类算法。

本文中, k-means、SC、US-ELM 和 US-EF-ELM 所有算法在数据集上运行 10 次, 记录平均结果和标准差。根据文献 [20] 在所有数据集上, US-ELM 隐藏节点数和 US-EF-ELM 隐藏节点数均设置为 2 000, US-ELM 和 US-EF-ELM 的激活函数均为 sigmoid 函数。在 SC-EF-ELM 中, 从 {100、200、500、1 000、2 000}中选择隐藏节点个数, 激活函数也为 sigmoid 函数。在 US-ELM、US-EF-ELM 和 SC-EF-ELM 中, 正则化参数选取范围为 $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$, 在 SC 和 SC-EF-ELM 中, 核函数为高斯核函数, 其中高斯核函数的参数为样本间的中值距离。

3.2　性能指标

假设数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, 通过聚类划分得到类簇 $\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k\}$, 真实的类划分为 $\boldsymbol{C}'=\{\boldsymbol{C}'_1,\boldsymbol{C}'_2,\cdots,\boldsymbol{C}'_k\}$, 同时, 令 $\boldsymbol{Y}$ 和 $\boldsymbol{Y}'$ 分别表示与 $\boldsymbol{C}$ 和 $\boldsymbol{C}'$ 对应的分配标签。实验中, 采用 F-measure (F_1)、聚类准确性 (cluster accuracy, ACC) 和标准化互信息 (normalized mutual information, NMI) 这 3 个评价标准来衡量聚类结果的质量及算法的有效性。对于这 3 个评价标准, 值越大, 聚类性能越好。F_1 表示精确率 (precision) 和召回率 (recall) 的加权调和平均值:

$$F=\frac{(a^2+1)PR}{a^2(P+R)}$$

当参数 $a=1$ 时, 就是 F_1 指标

$$F_1=\frac{2PR}{P+R}$$

式中: P 是精确率 (precision) 度量值; R 是召回率 (recall) 度量值。F_1 综合了 P 和 R 的结果, 当 F_1 较大时则比较说明实验结果比较理想。

ACC 表示聚类结果中被正确划分的数据点比例:

$$\mathrm{ACC}=\sum_{i}^{N}\delta(\boldsymbol{y}_i,\mathrm{map}(\boldsymbol{y}'_i))/n$$

式中: n 表示数据样本的个数; 当 $\boldsymbol{x}=\boldsymbol{y}$ 时, 则 $\delta(\boldsymbol{x},\boldsymbol{y})=1$, 否则 $\delta(\boldsymbol{x},\boldsymbol{y})=0$。map(·) 表示通过 Hungarian 算法将每个聚类标签映射到一个类标签, 并且映射是最佳。

NMI 衡量的是算法的划分质量:

$$\mathrm{NMI}=\frac{\sum_{i}^{k}\sum_{j}^{k'}|\boldsymbol{C}_i\cap\boldsymbol{C}'_j|\log\dfrac{n|\boldsymbol{C}_i\cap\boldsymbol{C}'_j|}{|\boldsymbol{C}_i||\boldsymbol{C}'_j|}}{\sqrt{\sum_{i=1}^{k}|\boldsymbol{C}_i|\log\dfrac{|\boldsymbol{C}_i|}{\boldsymbol{n}}\cdot\sum_{j=1}^{k'}|\boldsymbol{C}'_j|\log\dfrac{|\boldsymbol{C}'_j|}{n}}}$$

NMI 的值越大, 聚类有效性越好。

3.3　实验结果

所提出的 SC-ELM-AE 与 k-means、SC、US-ELM 和 US-EF-ELM 在 5 个数据集上的表现对比见表 2。

表 2　提出的 SC-ELM-AE 算法 5 个 UCI 数据集上的性能比较

Table 2　Performance comparison of the proposed SC-ELM-AE on five UCI datasets

数据集	评价指标	k-means	SC	US-ELM	US-EF-ELM	SC-ELM-AE
WDBC	F_1	0.8768±0.0000	0.8807±0.0000	0.7861±0.1464	0.8865±0.0014	**0.8887±0.0000**
	ACC	0.9279±0.0000	0.8807±0.0000	0.8356±0.1577	0.9338±0.0008	**0.9667±0.0000**
	NMI	0.6231±0.0000	0.6260±0.0000	0.4554±0.2542	0.6538±0.0047	**0.6460±0.0000**

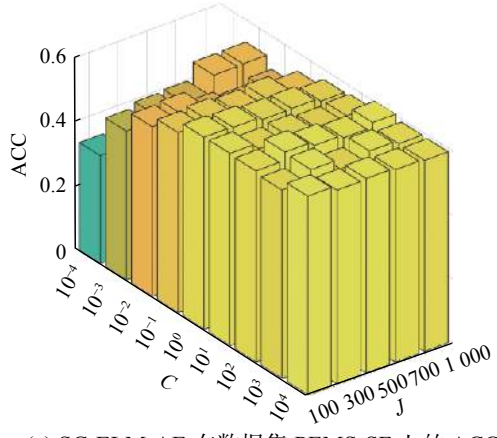
续表 2

数据集	评价指标	k-means	SC	US-ELM	US-EF-ELM	SC-ELM-AE
Ionosphere	F_1	0.604 3±0.009 7	0.597 6±0.000 0	0.664 4±0.056 0	0.658 5±0.053 3	**0.867 4±0.000 0**
	ACC	0.709 5±0.006 8	0.703 7±0.000 0	0.689 8±0.090 4	0.699 0±0.085 5	**0.875 1±0.000 0**
	NMI	0.130 0±0.012 5	0.126 4±0.000 0	0.152 9±0.106 4	0.153 2±0.090 6	**0.438 4±0.000 0**
Isolet	F_1	0.500 7±0.023 4	0.531 9±0.015 1	0.526 2±0.031 9	0.532 5±0.023 1	**0.644 4±0.018 5**
	ACC	0.532 6±0.030 8	0.564 9±0.023 8	0.570 1±0.036 2	0.583 0±0.032 0	**0.791 7±0.023 0**
	NMI	0.721 6±0.010 9	0.717 7±0.009 3	0.739 0±0.012 7	0.735 5±0.010 1	**0.745 3±0.008 7**
PEMS_SF	F_1	0.338 5±0.038 0	0.301 4±0.026 6	0.313 4±0.019 2	0.364 1±0.028 7	**0.436 5±0.020 4**
	ACC	0.312 4±0.018 5	0.316 9±0.022 2	0.329 8±0.023 5	0.377 6±0.037 1	**0.644 4±0.022 7**
	NMI	0.332 9±0.037 6	0.314 4±0.031 8	0.356 2±0.015 9	0.389 1±0.032 6	**0.475 6±0.011 0**
TDT2_10	F_1	0.364 1±0.028 7	0.338 9±0.009 3	0.738 3±0.086 5	0.735 6±0.107 6	**0.963 5±0.040 7**
	ACC	0.377 6±0.037 1	0.457 5±0.012 2	0.742 3±0.096 4	0.732 6±0.116 3	**0.968 4±0.043 7**
	NMI	0.389 1±0.032 6	0.511 4±0.011 6	0.857 2±0.041 8	0.860 9±0.050 0	**0.974 7±0.016 5**

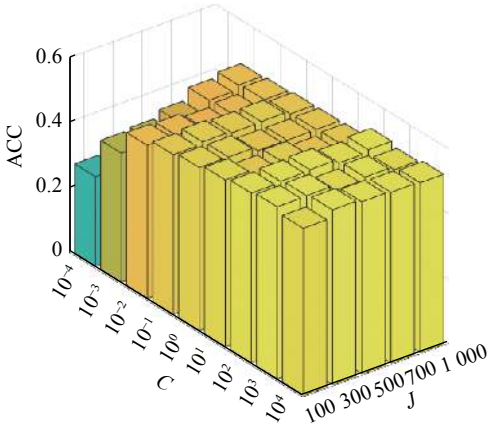
从表 2 中可以看出，本文所提出的 SC-ELM-AE 算法在聚类准确率上与对比算法相比在 5 个数据集上都有了较大提升。例如，对于高维数据集 PEMS-SF，本文所提算法与对比算法相比在 ACC 上分别提升了 33.2%、32.75%、31.46%、26.68% 平均提升了 31.02%。

SC-ELM-AE 聚类算法利用 ELM-AE 模型无需迭代即可获得低维特征表示空间，尽可能多地保留了原始数据集的丰富信息，使获得的聚类结果更加准确。实验数据结果表明，本文提出的 SC-ELM-AE 算法在进行实验的数据集上与对比算法相比聚类精度有较大的提升，这也验证了本文所提算法的合理性和有效性。

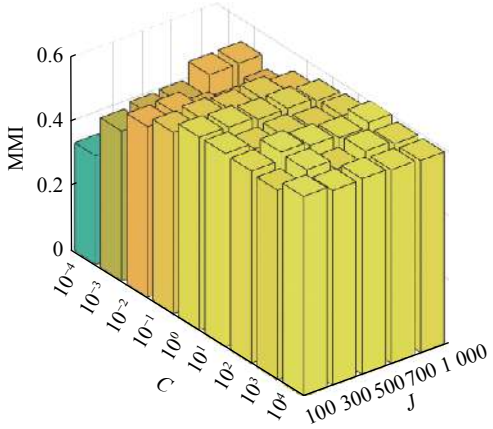
为了验证提出的 SC-ELM-AE 算法在增加隐含层节点后的表现，在实验中将 ELM-AE 模型结构的隐含层节点数从 100 增加到 2 000，并在数据集 WDBC 和 PEMS-SF 上基于 ELM-AE 的嵌入特征空间进行快速谱聚类，实验结果如图 2、3 所示。



(a) SC-ELM-AE 在数据集 PEMS-SF 上的 ACC



(b) SC-ELM-AE 在数据集 PEMS-SF 上的 F_1



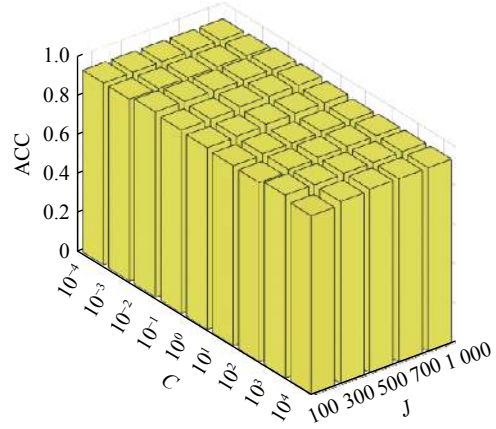
(c) SC-ELM-AE 在数据集 PEMS-SF 上的 NMI

图 2　在数据集 PEMS-SF 上 SC-ELM-AE 的性能变化

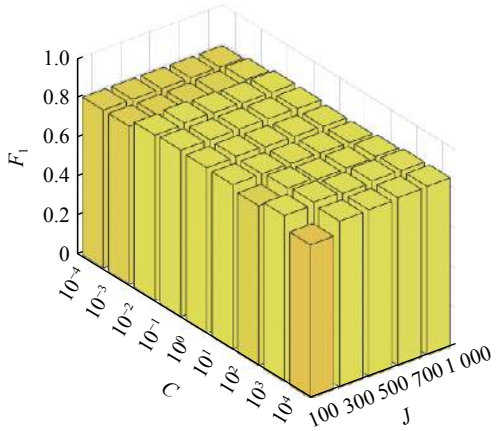
Fig. 2　Performances for SC-ELM-AE on dataset PEMS-SF

从图 2 可以看出：将 ELM-AE 特征表示空间作为谱聚类输入，从 $(10^{-4},10^{-3},10^{-2})$ 中选取正则化参数，当隐藏节点较小，ACC、F-measure 和 NMI 值较低。而在 $(10^{-1},10^{0},10^{1},10^{2},10^{3},10^{4})$ 选取正

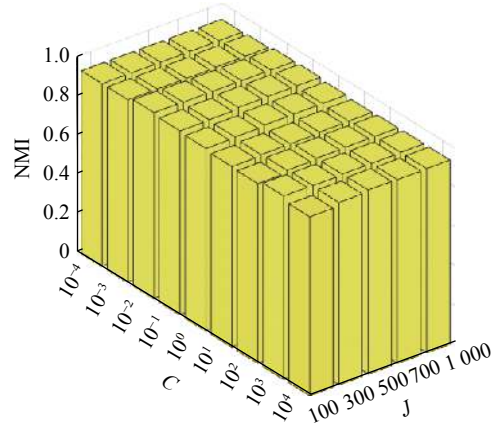
则化参数时, 隐藏节点数量的变化基本不会引起算法性能的波动。



(a) SC-ELM-AE 在数据集 WDBC 上的 ACC



(b) SC-ELM-AE 在数据集 WDBC 上的 F_1



(c) SC-ELM-AE 在数据集 WDBC 上的 NMI

图 3　在数据集 WDBC 上 SC-ELM-AE 的性能变化
Fig. 3　Performances for SC-ELM-AE on dataset WDBC

从图 3 可以看出, 本文提出的 SC-ELM-AE 始终是稳定的。因此, 可以推断参数的选择对算法性能的影响不大, 在合适的正则化参数下, 采用很少的隐藏节点即可实现较高的聚类精度, 与传统的聚类方法相比, 所提出的 SC-ELM-AE 算法具有更强的实用性。此外, 在 UCI 的其他 3 个基准数据集上进行验证, 实验结果与推断一致, 也证明了所提 SC-ELM-AE 的性能的有效性。

4　结束语

本文提出了一种通过 ELM-AE 特征表示空间的谱聚类算法。它利用 ELM-AE 将输入的原始数据集转化为数据特征表示空间, 再对特征表示空间样本集进行谱聚类, 利用 ELM-AE 获得的特征空间可以更好地反映出原始数据的主要信息且计算成本较低; 使用 ELM-AE 进行特征提取, 提高了聚类的准确性。通过实验验证了本文算法在有效性和准确性两方面优于现有的谱聚类算法, 能够快速有效地处理复杂高维数据。在未来的工作中需要考虑如何在保证聚类精确的情况下进一步提高聚类的速度以及对大规模数据的处理。

参考文献:

[1] BERKHIN P. A survey of clustering data mining techniques[M]//KOGAN J, NICHOLAS C, TEBOULLE M. Grouping Multidimensional Data. Berlin, Heidelberg: Springer, 2006: 25−71.

[2] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究 [J]. 软件学报, 2008, 19(1): 48–61.
SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research[J]. Journal of software, 2008, 19(1): 48–61.

[3] 刘兵. Web 数据挖掘 [M]. 俞勇, 薛贵荣, 韩定一, 译. 北京: 清华大学出版社, 2011.

[4] WU Junjie, LIU Hongfu, XIONG Hui, et al. K-means-based consensus clustering: a unified view[J]. IEEE transactions on knowledge and data engineering, 2015, 27(1): 155–169.

[5] WANG Yangtao, CHEN Lihui. Multi-view fuzzy clustering with minimax optimization for effective clustering of data from multiple sources[J]. Expert systems with applications, 2017, 72: 457–466.

[6] VAN LUXBURG U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395–416.

[7] JIA Hongjie, DING Shifei, XU Xinzheng, et al. The latest research progress on spectral clustering[J]. Neural computing and applications, 2014, 24(7/8): 1477–1486.

[8] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述 [J]. 计算机科学, 2008, 35(7): 14–18.
CAI Xiaoyan, DAI Guanzhong, YANG Libin. Survey on spectral clustering algorithms[J]. Computer science, 2008, 35(7): 14–18.

[9] HUANG Guangbin, CHEN Lei, SIEW C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE transactions

on neural networks, 2006, 17(4): 879–892.

[10] ZHANG Rui, LAN Yuan, HUANG Guangbin, et al. Universal approximation of extreme learning machine with adaptive growth of hidden nodes[J]. IEEE transactions on neural networks and learning systems, 2012, 23(2): 365–371.

[11] HUANG Guangbin, ZHOU Hongming, DING Xiaojian, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE transactions on systems, man, and cybernetics, part B (cybernetics), 2012, 42(2): 513–529.

[12] DA SILVA B L S, INABA F K, SALLES E O T, et al. Outlier Robust Extreme Machine Learning for multi-target regression[J]. Expert systems with applications, 2020, 140: 112877.

[13] ZENG Yijie, LI Yue, CHEN Jichao, et al. ELM embedded discriminative dictionary learning for image classification[J]. Neural networks, 2020, 123: 331–342.

[14] WU Chao, LI Yaqian, ZHAO Zhibiao, et al. Extreme learning machine with multi-structure and auto encoding receptive fields for image classification[J]. Multidimensional systems and signal processing, 2020, 31(4): 1277–1298.

[15] BARTLETT P L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network[J]. IEEE transactions on information theory, 1998, 44(2): 525–536.

[16] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504–507.

[17] BENGIO Y, YAO Li, ALAIN G, et al. Generalized denoising auto-encoders as generative models[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc., 2013: 899−907.

[18] BALDI P. Autoencoders, unsupervised learning and deep architectures[C]//Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop. Washington, USA: JMLR. org, 2011: 37−50.

[19] 袁非牛, 章琳, 史劲亭, 等. 自编码神经网络理论及应用综述 [J]. 计算机学报, 2019, 42(1): 203–230.
YUAN Feiniu, ZHANG Lin, SHI Jinting, et al. Theories and applications of auto-encoder neural networks: a literature survey[J]. Chinese journal of computers, 2019, 42(1): 203–230.

[20] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of machine learning research, 2010, 11(12): 3371–3408.

[21] 刘帅师, 程曦, 郭文燕, 等. 深度学习方法研究新进展 [J]. 智能系统学报, 2016, 11(5): 567–577.
LIU Shuaishi, CHENG Xi, GUO Wenyan, et al. Progress report on new research in deep learning[J]. CAAI Transactions on intelligent systems, 2016, 11(5): 567–577.

[22] 李建元, 周脚根, 关佶红, 等. 谱图聚类算法研究进展 [J]. 智能系统学报, 2011, 6(5): 405–414.
LI Jianyuan, ZHOU Jiaogen, GUAN Jihong, et al. A survey of clustering algorithms based on spectra of graphs[J]. CAAI transactions on intelligent systems, 2011, 6(5): 405–414.

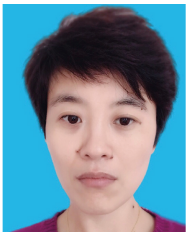
[23] FILIPPONE M, CAMASTRA F, MASULLI F, et al. A survey of kernel and spectral methods for clustering[J]. Pattern recognition, 2008, 41(1): 176–190.

[24] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[C]//Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, British Columbia, Canada: MIT Press, 2001: 849−856.

[25] KASUN L L C, ZHOU H, HUANG G B, et al. Representational learning with extreme learning machine for big data[J]. IEEE intelligent systems, 2013, 28(6): 31–34.

[26] DING Shifei, ZHANG Nan, ZHANG Jian, et al. Unsupervised extreme learning machine with representational features[J]. International journal of machine learning and cybernetics, 2017, 8(2): 587–595.

作者简介：



王丽娟，副教授，博士研究生，CCF 会员，主要研究方向为机器学习、聚类分析。

丁世飞，教授，博士生导师，博士，CCF 杰出会员，第八届吴文俊人工智能科学技术奖获得者，主要研究方向为人工智能与模式识别，机器学习与数据挖掘。主持国家重点基础研究计划课题 1 项、国家自然科学基金面上项目 3 项。出版专著 5 部，发表学术论文 200 余篇。