文章编号：2095-4980(2020)03-0497-07

数据挖掘中一种改进的谱组合聚类算法

童绪军，吴义春

(安徽医学高等专科学校 基础部，安徽 合肥 230601)

摘　要：组合聚类(EC)是解决数据挖掘问题的关键手段之一，但现有的 EC 方法较少考虑可能破坏聚类结构的各种噪声，降低了聚类性能。为此，提出一种改进的谱组合聚类(ISEC)方法。将聚类问题建模为输入的多个基本分区(BPs)派生的共协矩阵的图分割问题；ISEC 方法学习得到共协矩阵的低秩表示，并在共协矩阵上进行谱聚类，提高聚类性能；最后采用增强拉格朗日乘数法进行优化求解，获得最终的聚类结果。在多个真实数据集上的仿真实验结果表明，ISEC 方法的聚类性能优于目前的大多数聚类方法。

关键词：组合聚类；基本分区；低秩表示；共协矩阵；增强拉格朗日乘数法

中图分类号：TN911.7　　**文献标志码：**A　　**doi：**10.11805/TKYDA2019338

An improved spectral Ensemble Clustering algorithm in data mining

TONG Xujun，WU Yichun

(Basic Department，Anhui Medical College，Hefei Anhui 230601，China)

Abstract：Ensemble Clustering(EC) is one of the key means to solve data mining problems, but the existing EC methods rarely consider the various noises that may damage the clustering structure and reduce the clustering performance. To solve this problem, an Improved Spectral Ensemble Clustering(ISEC) method is proposed. Firstly, the clustering problem is modeled as a graph partitioning problem of coincidence matrices derived from inputting multiple Basic Partitions(BPs). Then, The ISEC method learns to obtain the low rank representation of the covariance matrix, and carries on the spectral clustering to improve the clustering performance. Finally, the optimization solution is carried out by the enhanced Lagrange multiplier method, so as to obtain the final clustering result. The simulation results on several real data sets show that the clustering performance of ISEC method is better than that of most existing clustering methods.

Keywords：Ensemble Clustering；Basic Partitions；Low Rank Representation(LRR)；covariance matrix；enhanced Lagrange multiplier method

在数据挖掘、信息检索、机器学习和计算机视觉等领域，聚类分析[1-2]是一项核心技术。已有大量研究基于不同的假设(如连通性、形心、分布、密度和子空间等)提出了多个聚类算法。由于不同的算法会获得不同的效果，因此很难在实践中确定使用哪种算法。近年来，组合聚类(EC)[3]算法引起了极大关注，并成为传统聚类方法[4]的强大替代性方法。该方法的输入通常为一组基本分区(BPs)，旨在将多个 BPs 融合为一个共有分区。在设计 EC 算法的过程中，BPs 的生成和聚合为 2 个关键步骤[5]。一般可通过以下 3 种策略生成多个 BPs：a) 在同一个数据集上执行多个不同的基本聚类算法[6]；b) 在来自同一个数据集的多个子样本上生成 BPs[7]；c) 设置不同的参数来多次运行相同的聚类方法，如随机参数选择方法[8]，该方法被视为最成功的 BPs 生成方法。

现有的 EC 算法可分为两类[9]：基于效用函数的方法和基于共协矩阵的方法。第一类方法首先采用效用函数测量分区和多个 BPs 之间的相似性，然后通过目标函数的最大化找到最终分区。如，WU 等[10]采用 KCC(K-means method based on Common Clustering)效用函数将组合聚类转化为 K 均值聚类问题进行求解。针对组合聚类问题，相关文献还提出了一些典型方法，如基于非负矩阵因子分解的方法[11]、基于核的方法[12]和模拟退火法等。

收稿日期：2019-09-11；**修回日期：**2019-11-04

基金项目：安徽省自然科学基金资助项目(2017jyxm0606；2017jyxm0608；2019xjzr01)

作者简介：童绪军(1979–)，男，硕士，讲师，主要研究方向为数据挖掘、大数据处理等。email:1835428764@qq.com

第二类方法将输入的 BPs 集成在一个共协矩阵中，并将 EC 转化为图分割问题。然而大多数现有基于共协矩阵的聚类方法未考虑来自输入 BPs 的各种噪声，它们可能会严重破坏共协矩阵的聚类结构，导致 EC 方法产生误差，影响聚类的最终性能。很少的文献[13-14]研究了该问题，其中，文献[13]将 BPs 之间的不一致视为多个不确定数据对进行处理，标为缺失值，并通过矩阵补全将共协矩阵作为低秩矩阵进行恢复，但实际上很难决定数据的不确定度；文献[14]试图学习 BPs 中存在的异常值，并通过最小化异常值和低秩约束下每个输入 BP 之间的 KL(Kullback-Leibler)散度获得一个稳健的共协矩阵。但随着 BPs 数量的增加，该方法会产生很高的空间复杂度，限制它在大规模数据集中的适用性。此外，这两类方法的学习过程缺乏来自聚类任务的明确指导，将共协矩阵学习和聚类过程看作 2 个单独的步骤，未能提供一个统一框架。

为解决上述问题，本文提出一种新的谱组合聚类(ISEC)方法。该方法首先学习得到共协矩阵 $\boldsymbol{S}$ 的低秩表示 $\boldsymbol{Z}$，并通过 $\boldsymbol{L}_Z$ 上的谱聚类[15]找到共有分区 $\boldsymbol{H}$。利用 $\boldsymbol{Z}$ 揭示 $\boldsymbol{S}$ 的聚类结构，并通过稀疏误差矩阵 $\boldsymbol{E}$ 捕捉 $\boldsymbol{S}$ 内的噪声。在学习过程中，利用 $\boldsymbol{H}$ 迭代提高 $\boldsymbol{Z}$ 的区块对角化结构，并从 $\boldsymbol{H}$ 或 $\boldsymbol{Z}$ 中获得最终聚类结果。

1 低秩矩阵分析

低秩矩阵分析作为一种从包含误差(如噪声、缺失项、损坏和异常值)的样本中恢复数据的有效工具，已得到广泛应用。它有 2 种表示形式[16]：稳健主成分分析(Principal Component Analysis，PCA)和低秩表示(LRR)。给定一个矩阵 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]\in\mathbf{R}^{d\times n}$，其中，每个列向量 $\boldsymbol{x}_i\in\mathbf{R}^d$ 表示一个样本，稳健 PCA 将 $\boldsymbol{X}$ 分解为一个低秩矩阵 $\boldsymbol{A}$ 和一个稀疏矩阵，分别用来恢复数据和确定误差。该方法采用来自单个子空间的数据进行低秩矩阵恢复和补全任务。而 LRR 假设从多个低维度联合子空间提取数据，并通过 $\boldsymbol{X}$ 的最低秩表示 $\boldsymbol{Z}$ 恢复这些子空间：

$$\begin{cases}\min\limits_{\boldsymbol{Z},\boldsymbol{E}} rank(\boldsymbol{Z})+\lambda\|\boldsymbol{E}\|_0\\ \text{s.t. } \boldsymbol{X}=\boldsymbol{X}\boldsymbol{Z}+\boldsymbol{E}\end{cases} \tag{1}$$

式中 $\lambda>0$，用于平衡 $\boldsymbol{Z}$ 的秩和误差矩阵 $\boldsymbol{E}$ 的稀疏性。

式(1)属于 NP 难问题[16]，通常使用核范数估计秩 $\boldsymbol{Z}$，使用 l_1 或 $l_{2,1}$ 范数估计$\|\boldsymbol{E}\|_0$。值得注意的是，对式(1)求极小值也可以获得 $\boldsymbol{X}$ 的低秩矩阵 $\boldsymbol{XZ}$，LRR 可视为稳健 PCA 的泛化。$\boldsymbol{Z}$ 是一个相似度矩阵，可揭示数据点之间的隶属度，具有良好的分块对角特性[16]，可发现数据的全局结构，用于协助聚类任务。因此，本文采用 LRR 学习共协矩阵的稳健表示，而非直接将其恢复为低秩矩阵。

2 谱组合聚类

2.1 组合聚类

设 $X=\{x_1,x_2,\cdots,x_n\}$为从 K 个聚类($C=\{C_1,C_2,\cdots,C_k\}$)中采样得到的 n 个数据点的集合。$\Pi=\{\pi_1,\pi_2,\cdots,\pi_r\}$表示输入的 r 个基本分区(BPs)，每个基本分区将 X 中的数据点分成 K_i 个明确的分区，即将数据点映射到集合 $\pi_i=\{\pi_i(x_1),\pi_i(x_2),\cdots,\pi_i(x_n)\}$ 中。其中，K_i 为第 i 个 BP 的聚类数量，$1\leqslant\pi_i(x_j)\leqslant K_i$，且 $1\leqslant i\leqslant r,1\leqslant j\leqslant n$。组合聚类的目标是发现和输入 BPs 最一致的共有分区，并将 X 分配到原始的 K 个聚类中。通常，EC 方法可将 r 个 BPs 集成为一个共协矩阵 $\boldsymbol{S}\in\mathbf{R}^{n\times n}$，然后通过计算 2 个数据点出现在同一个聚类的次数获得最终共有聚类 π：

$$S(x_p,x_q)=\sum_{i=1}^{r}\delta\left[\pi_i(x_p),\pi_i(x_q)\right] \tag{2}$$

式中：如果 $a=b$，则 $x_p,x_q\in X$，且 $\delta(a,b)=1$；否则为 0。显然，应通过 $\boldsymbol{S}=\boldsymbol{S}/r$ 归一化 $\boldsymbol{S}$。根据文献[16]，在共协矩阵 $\boldsymbol{S}$ 上进行谱聚类，并通过式(3)获得其阵迹最小化形式：

$$\begin{cases}\min\limits_{\boldsymbol{H}} tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}_S\boldsymbol{H})\\ \text{s.t.}\,\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}\end{cases} \tag{3}$$

式中：$\boldsymbol{L}_S=\boldsymbol{I}-\boldsymbol{D}_S^{-1/2}\boldsymbol{S}\boldsymbol{D}_S^{-1/2}$ 为 $\boldsymbol{S}$ 的归一化拉普拉斯矩阵，$\boldsymbol{D}_S\in\mathbf{R}^{n\times n}$ 为度矩阵，是对角矩阵，其第 j 个对角元素为 $\boldsymbol{S}$ 第 j 列的和，而 $\boldsymbol{H}\in\mathbf{R}^{n\times K}$ 定义为 π 的扩展分区矩阵：

$$H_{jk}=\begin{cases}1\Big/\sqrt{|C_k|}, & \text{if } x_j\in C_k \text{ in } \pi\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

使用 $\boldsymbol{H}$ 表示最终的组合聚类结果。

2.2 问题描述

根据式(3)可知，共协矩阵 $\boldsymbol{S}$ 的固有结构是 EC 方法的关键因素。理想情况下，$\boldsymbol{S}$ 应为低秩矩阵($K<<n$)，并具有明确的 K 区块对角化聚类结构。通常使用多个 BPs 直接计算 $\boldsymbol{S}$，BPs 中存在的噪声会轻易破坏其聚类结构。LRR 可用来处理噪声，并通过分块对角形式揭示数据点之间的隶属度，因此，可使用 LRR 学习共协矩阵的稳健表示。给定归一化共协矩阵 $\boldsymbol{S}$，ISEC 的目标函数为：

$$\begin{cases}\min\limits_{\boldsymbol{H},\boldsymbol{Z},\boldsymbol{E}} tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}_Z\boldsymbol{H})+\lambda_1\|\boldsymbol{Z}\|_*+\lambda_2\|\boldsymbol{E}\|_{2,1}\\ \text{s.t. } \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I},\boldsymbol{S}=\boldsymbol{S}\boldsymbol{Z}+\boldsymbol{E}\end{cases} \tag{5}$$

式中：$\lambda_1,\lambda_2>0$ 为 2 个惩罚因子；$\boldsymbol{L}_Z$ 为通过 $\boldsymbol{Z}$ 和 $\boldsymbol{H}$ 构造的图形的归一化拉普拉斯矩阵：

$$\boldsymbol{L}_Z=\boldsymbol{I}-\boldsymbol{D}_Z^{-1/2}\left[(\boldsymbol{Z}+\boldsymbol{Z}^{\mathrm{T}})/2+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right]\boldsymbol{D}_Z^{-1/2} \tag{6}$$

通过式(7)计算度矩阵 $\boldsymbol{D}_Z$：

$$\boldsymbol{D}_Z=diag\left([d_1,d_2,\cdots,d_n]\right) \tag{7}$$

式中 $d_j(1\leqslant j\leqslant n)$为矩阵 $(\boldsymbol{Z}+\boldsymbol{Z}^{\mathrm{T}})/2+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$ 第 j 列的和。

2.3 优化求解

式(5)不是关于 $\boldsymbol{Z}$ 和 $\boldsymbol{H}$ 的凸问题，很难直接求解。本文将其分解为几个子问题，通过固定其他变量对每个子问题进行优化，并采用增强拉格朗日乘数法(Augmented Lagrange Multiplier，ALM)[17]解决该问题。首先引入辅助变量 $\boldsymbol{J}$ 对式(5)进行拆分，并将其等价转换为：

$$\begin{cases}\min\limits_{\boldsymbol{H},\boldsymbol{Z},\boldsymbol{E}} tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}_Z\boldsymbol{H})+\lambda_1\|\boldsymbol{J}\|_*+\lambda_2\|\boldsymbol{E}\|_{2,1}\\ \text{s.t. } \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I},\boldsymbol{S}=\boldsymbol{S}\boldsymbol{Z}+\boldsymbol{E},\boldsymbol{Z}=\boldsymbol{J}\end{cases} \tag{8}$$

参照文献[16]对约束条件 $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}$ 进行松弛，可得式(8)的增广拉格朗日函数为：

$$\boldsymbol{L}=tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}_Z\boldsymbol{H})+\lambda_1\|\boldsymbol{J}\|_*+\lambda_2\|\boldsymbol{E}\|_{2,1}+<Y_1,\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}-\boldsymbol{E}>+<Y_2,\boldsymbol{Z}-\boldsymbol{J}>+\frac{\mu}{2}\left(\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}-\boldsymbol{E}\|_F^2+\|\boldsymbol{Z}-\boldsymbol{J}\|_F^2\right) \tag{9}$$

式中：Y_1,Y_2 为 2 个拉格朗日乘数；$\mu>0$ 为惩罚因子。ALM 求解程序通过迭代更新的方法解出式(9)，按次序处理 $\boldsymbol{J},\boldsymbol{Z},\boldsymbol{E}$ 和 $\boldsymbol{H}$，并在其他变量固定的情况下，一次优化一个变量。具体过程如下：

1) 更新 $\boldsymbol{J}$：首先，根据 $\boldsymbol{J}$ 最小化 $\boldsymbol{L}$，并通过式(10)获得 $\boldsymbol{J}^{(t+1)}$：

$$\arg\min_{\boldsymbol{J}}\lambda_1\|\boldsymbol{J}\|_*+<Y_2^{(t)},\boldsymbol{Z}^{(t)}-\boldsymbol{J}>+\frac{\mu^{(t)}}{2}\left\|\boldsymbol{Z}^{(t)}-\boldsymbol{J}\right\|_F^2=\arg\min_{\boldsymbol{J}}\frac{\lambda_1}{\mu^{(t)}}\|\boldsymbol{J}\|_*+\frac{1}{2}\left\|\boldsymbol{J}-\left(\boldsymbol{Z}^{(t)}+\frac{1}{\mu^{(t)}}Y_2^{(t)}\right)\right\|_F^2 \tag{10}$$

通过奇异值阈值(Singular Value Thresholding，SVT)算子[18]解出式(10)，得到以下形式的闭式解：

$$\boldsymbol{J}^{(t+1)}=\Theta_{\frac{\lambda_1}{\mu^{(t)}}}\left(\boldsymbol{Z}^{(t)}+\frac{1}{\mu^{(T)}}Y_2^{(t)}\right) \tag{11}$$

式中 $\Theta(\cdot)$为 SVT 算子。

2) 更新 $\boldsymbol{Z}$：把式(6)代入式(9)中的 $\boldsymbol{L}$，去掉不相关项，更新 $\boldsymbol{Z}$ 的子问题等价于：

$$\begin{aligned}\arg\min_{\boldsymbol{Z}}-\frac{1}{2}tr\left[\boldsymbol{H}^{(t)\mathrm{T}}\boldsymbol{D}_Z^{-1/2}(\boldsymbol{Z}+\boldsymbol{Z}^{\mathrm{T}})\boldsymbol{D}_Z^{-1/2}\boldsymbol{H}^{(t)}\right]+<Y_1^{(t)},\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}-\boldsymbol{E}^{(t)}>+<Y_1^{(t)},\boldsymbol{Z}-\boldsymbol{J}^{(t+1)}>+\\ \frac{\mu^{(t)}}{2}\left(\left\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}-\boldsymbol{E}^{(t)}\right\|_F^2+\left\|\boldsymbol{Z}-\boldsymbol{J}^{(t+1)}\right\|_F^2\right)\end{aligned} \tag{12}$$

其中，对 $\boldsymbol{Z}$ 的求导 $\boldsymbol{D}_Z$ 相对复杂，增加了求解 $\boldsymbol{Z}^{(t+1)}$的难度。为简化解，可将 $\boldsymbol{D}_Z$ 固定为常数，通过定义 $\boldsymbol{Z}^{(t+1)}$和 $\boldsymbol{H}^{(t)}$ 进行更新。通过固定 $\boldsymbol{D}_Z$，式(12)成为 $\boldsymbol{Z}$ 的二次规划问题。因此，通过求 $\boldsymbol{L}$ 关于 $\boldsymbol{Z}$ 的导数，可以获得 $\boldsymbol{Z}^{(t+1)}$：

$$\boldsymbol{Z}^{(t+1)}=(\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}+\boldsymbol{I})^{-1}\left[\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}+\boldsymbol{J}^{(t+1)}-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{E}^{(t)}+\frac{1}{\mu^{(t)}}\left(\boldsymbol{S}^{\mathrm{T}}Y_1^{(t)}-Y_2^{(t)}+\boldsymbol{D}_Z^{-1/2}\boldsymbol{H}^{(t)}\boldsymbol{H}^{(t)\mathrm{T}}\boldsymbol{D}_Z^{-1/2}\right)\right] \tag{13}$$

3) 更新 $\boldsymbol{E}$：更新 $\boldsymbol{E}$ 的子问题可以转换为：

$$\arg\min_{\boldsymbol{E}}\lambda_2\|\boldsymbol{E}\|_{2,1}+<Y_1^{(t)},\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}^{(t+1)}-\boldsymbol{E}>+\frac{\mu^{(t)}}{2}\left(\left\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}^{(t+1)}-\boldsymbol{E}\right\|_F^2\right)=\arg\min_{E}\frac{\lambda_2}{\mu^{(t)}}\|\boldsymbol{E}\|_{2,1}+\frac{1}{2}\left\|\boldsymbol{E}-\left(\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}^{(t+1)}+\frac{Y_1^{(t)}}{\mu^{(t)}}\right)\right\|_F^2 \tag{14}$$

根据文献[19]中的引理，可以按列解决该问题，其中，$\boldsymbol{E}^{(t+1)}$的每列具有闭式解：

$$\forall i\ \boldsymbol{E}_i^{(t+1)}=\begin{cases}\dfrac{\|\boldsymbol{Q}_i\|_2-\lambda_2/\mu^{(t)}}{\|\boldsymbol{Q}_i\|_2}, & \text{if } \|\boldsymbol{Q}_i\|_2>\lambda_2/\mu^{(t)}\\ 0, & \text{otherwise}\end{cases} \tag{15}$$

式中$\|\boldsymbol{Q}_i\|_2$为一个列向量的 l_2 范数，$1\leqslant i\leqslant n$，并且 $\boldsymbol{Q}=\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}^{(t+1)}+\dfrac{Y_1^{(t)}}{\mu^{(t)}}$。

4) 更新 $\boldsymbol{H}$：为解决 $\boldsymbol{H}$ 子问题，首先通过式(6)~(7)中 $\boldsymbol{Z}^{(t+1)}$和 $\boldsymbol{H}^{(t)}$的定义来更新 $\boldsymbol{L}_Z$ 和 $\boldsymbol{D}_Z$。可通过式(16)获得 $\boldsymbol{H}^{(t+1)}$：

$$\begin{cases}\arg\min\limits_{\boldsymbol{H}} tr(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}_Z\boldsymbol{H})\\ \boldsymbol{L}_Z=\boldsymbol{I}-\boldsymbol{D}_Z^{-1/2}\left[(\boldsymbol{Z}^{(t+1)}+\boldsymbol{Z}^{(t+1)\mathrm{T}})/2+\boldsymbol{H}^{(t)}\boldsymbol{H}^{(t)\mathrm{T}}\right]\boldsymbol{D}_Z^{-1/2}\end{cases} \tag{16}$$

更新乘数：$\boldsymbol{J}^{(t+1)}$,$\boldsymbol{Z}^{(t+1)}$,$\boldsymbol{E}^{(t+1)}$和 $\boldsymbol{H}^{(t+1)}$固定不变，通过步长为 $\mu^{(t)}$的梯度上升来计算拉格朗日乘数：

$$\begin{cases}Y_1^{(t+1)}=Y_1^{(t)}+\mu^{(t)}(\boldsymbol{S}-\boldsymbol{S}\boldsymbol{Z}^{(t+1)}-\boldsymbol{E}^{(t+1)})\\ Y_2^{(t+1)}=Y_2^{(t)}+\mu^{(t)}(\boldsymbol{Z}^{(t+1)}-\boldsymbol{J}^{(t+1)})\end{cases} \tag{17}$$

当求解式(5)的过程终止时，可最终获得稳健表示 $\boldsymbol{Z}$ 和优化分区矩阵 $\boldsymbol{H}$。此时，可以通过 $\boldsymbol{H}$ 上的 K 均值或 $\boldsymbol{Z}$ 上的谱聚类获得最终分区。

2.4 复杂度分析

ISEC 算法的计算成本主要包括 4 个部分：a) 为更新 $\boldsymbol{J}$，式(11)中的 SVD 计算开销为 $O(n^3)$，当 n 很大时，计算成本会很高。但可通过 SVD 分解加速为 $O(mn^2)$，其中，$m<<n$ 为矩阵 $\boldsymbol{J}$ 的秩。b) 式(13)中涉及多个矩阵相乘和一个矩阵逆运算，因此计算开销为 $O[(l+1)n^3]$，其中，l 为相乘的次数。c) 通过式(15)中的阈值策略更新 $\boldsymbol{E}$，其复杂度为 $O(n^2)$。d) 为了找到分区 $\boldsymbol{H}$，进行特征值分解，其复杂度为 $O(n^2)$。因此，ISEC 算法的总成本为 $O[t(m+1)n^2+(l+2)n^3]$，其中，t 为 ISEC 算法的迭代次数。

3 仿真实验

3.1 实验设置

实验中，采用 12 个真实数据集评估所提出 ISEC 算法的聚类性能。分别从 UCI 机器学习数据库(https://archive.ics.uci.edu/ml/datas-ets.html)选择 5 个数据集，从 CLUTO 数据库(http://glaros.dtc.umn.edu/gkhome/cluto/cluto/down-load)选择 4 个文本类数据集。另外，为进行综合评估，使用了来自其他来源的 3 个图像类数据集，COIL20[20], Dslr(https://www.eecs.berke-ley.edu/~jhoffman/domainadapt/)和 MNIST4K(htt-p://www.cad.zju.edu.cn/home/dengcai/Data/MLDat-a.html)。表 1 为所有数据集的更多详细信息。

表1 数据集详细信息

Table1 Details of data set

dataset	#instance	#feature	#class	source
breast_w	699	9	2	UCI
iris	150	4	3	UCI
ionosphere	351	35	2	UCI
pendigits	10 992	16	10	UCI
wine	178	13	3	UCI
fbis	2 463	2 000	17	CLUTO
re0	1 504	2 886	13	CLUTO
tr12	313	5 804	8	CLUTO
wap	1 560	846	20	CLUTO
COIL20	1 440	1 024	20	others
Dslr	157	800	10	others
MNIST4K	4 000	748	10	others

将本文提出的 ISEC 方法和目前较为经典和具有先进性能的几种组合聚类方法进行比较，包括：基于图形的共有聚类(Co-Clustering based on Graph，GCC)[3]、共协矩阵的分层聚类(Hierarchical Clustering of Covariance matrix，HCC)[4]、基于矩阵补全的组合聚类(Ensemble Clustering based on Matrix Completion，ECMC)[5]、基于共有聚类的 K 均值方法(KCC)[10]、谱组合聚类(Spectral Ensemble Clustering，SEC)[21]和稳健聚类组合(Robust Clustering Ensemble，RCE)[14]。采用 K 均值和谱聚类算法作为基准方法。

检验标准：采用目前较为典型的聚类检验标准对本文算法进行定量分析：

1) 平均聚类精确度(Average Clustering Accuracy，ACC)。给定一个包含 n 个实例 K 聚类的数据集 X，ACC 的计算公式为[22]：

$$\max_f \frac{1}{n}\sum_{j=1}^{n}\delta\left[y_j,f\left(\pi(x_j)\right)\right] \tag{18}$$

式中：$x_j\in X$, $y_j\in[1,K]$为 x_j 的实际聚类，$1\leqslant j\leqslant n$；π 为聚类结果。

2) 归一化互信息(Normalized Mutual Information，NMI)。它测量所产生的标签和实际标签之间的交互信息熵，定义为[23]：

$$NMI=\frac{\sum_h\sum_l n_{h,l}\log\frac{nn_{h,l}}{n_h n_l}}{\sqrt{\left(\sum_h n_h\log\frac{n_h}{n}\right)\left(\sum_l n_l\log\frac{n_l}{n}\right)}} \tag{19}$$

式中：n_h 和 n_l 分别为分区中发现的聚类 C_{h} 中的实例数量和实际聚类 C_l 中的实例数量；$n_{h,l}$ 为 C_{h} 和 C_l 中的实例数量。如果数据是随机分区的，NMI 会趋向于 0。ACC 和 NMI 的取值范围均是 0~1，值越高，说明聚类性能越好。

对于每个数据集，通过随机参数选择方法生成 r=100 个 BPs(记为 II)，并将 II 作为所有 EC 方法的默认输入。通过运行 K 均值算法获得 II 的每个 BP，聚类数量从 K 变化到 $\sqrt{n}$，其中，K 为实际聚类数量，n 为数据集大小。在执行每个 EC 方法时，直接运行作者的代码，并根据其论文的说明对参数进行微调。在传统方法中，通过 Matlab 函数 K-means 直接获得 K 均值，并根据文献[15]进行谱聚类。将聚类数量设为 K，以测试所有方法。另外，由于 KCC,SEC,K 均值和谱聚类均涉及随机初始化，进行 20 次试验，并报告平均结果。对于 ISEC 方法，将 λ_1=0.1 和 λ_2=0.01 设为默认设置。在配备 Intel Core i7 3.4 GHz CPU 和 32 GB RAM 的 64 位 Ubuntu 14.04 平台上，通过 Matlab 进行所有实验。需注意的是，RCE 存在高空间复杂度 $O(m^2)$，由于内存限制，在一些数据集上无法运行该方法，在表 2 和表 3 中标为“—”。

表2 不同方法在12个实际数据集上的聚类性能(ACC)

Table2 Clustering performance of different methods on 12 real datasets(ACC)

datasets	ISEC	ensemble clustering						baseline methods	
		GCC	HCC	ECMC	KCC	SEC	RCE	K-means	spectral
breast_w	0.971 5	0.905 6	0.949 9	0.958 5	0.663 5	0.958 1	0.971 5	0.957 5	0.962 4
iris	0.973 3	0.973 3	0.893 3	0.886 7	0.889 2	0.960 0	0.900 0	0.892 7	0.886 7
ionosphere	0.675 2	0.671 8	0.683 8	0.572 6	0.638 2	0.635 3	0.675 2	0.712 3	0.703 7
pendigits	0.864 4	0.730 6	0.742 4	0.783 0	0.639 7	0.746 1	—	0.745 1	0.707 5
wine	0.522 5	0.516 9	0.500 0	0.533 7	0.504 9	0.511 2	0.500 0	0.500 0	0.511 2
fbis	0.570 0	0.454 9	0.542 8	0.495 7	0.489 7	0.483 7	0.534 7	0.286 2	0.193 3
re0	0.438 2	0.355 1	0.395 6	0.423 5	0.362 7	0.349 1	0.354 4	0.372 0	0.291 1
tr12	0.693 3	0.555 9	0.450 5	0.476 0	0.560 3	0.588 5	0.501 6	0.282 0	0.288 5
wap	0.496 8	0.420 4	0.414 7	0.465 4	0.420 8	0.387 6	0.408 8	0.370 2	0.387 3
COIL20	0.666 0	0.536 8	0.304 9	0.520 8	0.611 8	0.610 3	0.407 7	0.630 9	0.640 4
Dslr	0.566 9	0.554 1	0.528 7	0.503 2	0.528 5	0.526 4	0.527 8	0.381 9	0.431 1
MINISTSK	0.613 3	0.660 8	0.582 0	0.515 5	0.557 5	0.570 4	—	0.543 2	0.528 7

表3 不同方法在12个实际数据集上的聚类性能(NMI)

Table3 Clustering performance of different methods on 12 real datasets(NMI)

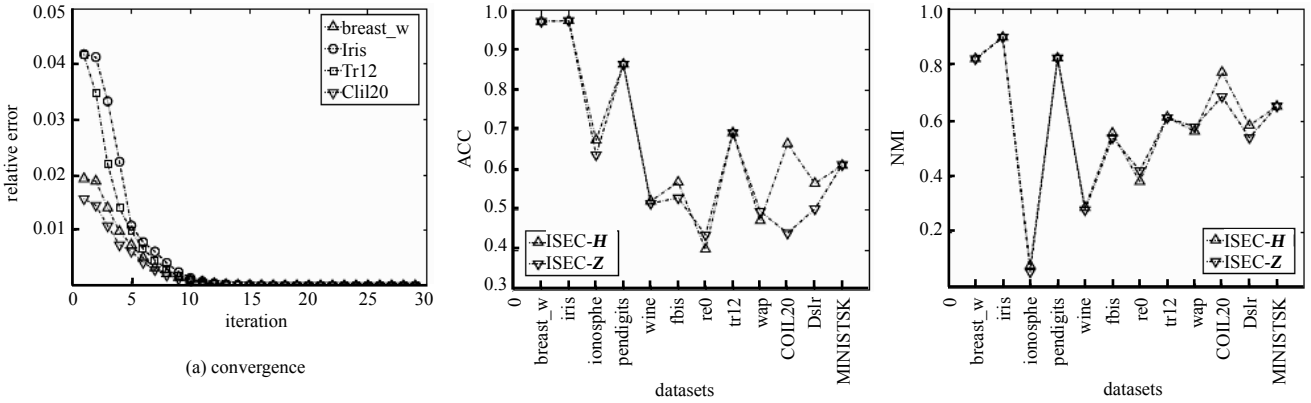
datasets	ISEC	ensemble clustering						baseline methods	
		GCC	HCC	ECMC	KCC	SEC	RCE	K-means	spectral
breast_w	0.823 8	0.614 4	0.699 9	0.736 1	0.281 6	0.736 1	0.806 4	0.736 1	0.756 3
iris	0.901 1	0.901 1	0.790 8	0.741 9	0.783 0	0.870 5	0.786 9	0.756 6	0.741 9
ionosphere	0.080 1	0.071 9	0.085 7	0.078 8	0.088 1	0.077 0	0.073 5	0.134 9	0.126 4
pendigits	0.826 4	0.705 8	0.772 9	0.734 4	0.683 6	0.738 7	—	0.689 3	0.659 4
wine	0.288 9	0.163 2	0.186 7	0.229 2	0.165 1	0.176 0	0.163 4	0.133 8	0.133 6
fbis	0.556 5	0.542 3	0.562 1	0.546 5	0.551 1	0.548 7	0.541 7	0.247 5	0.050 2
re0	0.422 1	0.380 2	0.377 1	0.404 6	0.390 4	0.382 4	0.326 4	0.232 4	0.283 9
tr12	0.613 8	0.490 3	0.418 9	0.396 3	0.512 1	0.513 9	0.457 2	0.089 0	0.073 5
wap	0.579 2	0.498 1	0.565 9	0.503 4	0.564 0	0.541 0	0.549 7	0.425 5	0.491 0
COIL20	0.775 4	0.714 1	0.612 4	0.685 5	0.756 3	0.750 0	0.619 3	0.767 0	0.769 8
Dslr	0.584 0	0.581 6	0.563 6	0.513 2	0.563 9	0.568 1	0.571 8	0.409 2	0.416 6
MINISTSK	0.653 4	0.599 8	0.622 7	0.531 5	0.564 7	0.565 2	—	0.453 7	0.455 6

3.2 不同算法的聚类性能比较

从表 2 可以看出，大多数情况下，ISEC 方法的性能优于其他方法。在 12 个数据集中，ISEC 方法在其中 9 个数据集表现出最优性能，在剩余数据集中的 ACC 性能次优。从表 3 可以看出，ISEC 方法在 12 个数据集的 10 个数据集中取得最优 NMI 性能，在其他方面获得次优性能。在数据集 breast_w,iris,pendigits,re0,tr12,wap,COIL20 和 Dslr 上，可以很明显看出本文方法较其他方法的优越性，其中，ISEC 在这 8 个数据集上同时取得最优 ACC 和 NMI 性能。

3.3 ISEC算法讨论

通过$||\boldsymbol{S}-\boldsymbol{SZ}-\boldsymbol{E}||_F/||\boldsymbol{S}||_F$计算相对误差，以评估ISEC方法的收敛性。如图1(a)所示，在4个数据集上，ISEC方法在 15 次迭代内收敛，这表明本文方法具有快速稳定的收敛性能。根据 ISEC 算法可知，ISEC 可通过在最优分区矩阵 $\boldsymbol{H}$(ISEC-$\boldsymbol{H}$)上运行 K 均值或在学习得到表示 $\boldsymbol{Z}$(ISEC-$\boldsymbol{Z}$)上进行谱聚类获得最终聚类结果。图 1(b)和图 1(c)展示了这 2 种方法之间的差异。大多数情况下，ISEC-$\boldsymbol{H}$ 和 ISEC-$\boldsymbol{Z}$ 具有类似性能。图 2 为在 tr12 和 wap 数据集上，λ_1 和 λ_2 这 2 个参数对 ISEC 性能(ACC 和 NMI)的影响。其中，参数范围均在 10^{-4}~1 之间。可以看出，在图 2 所示 4 种不同情况下，在类似区域[0.01, 1]内，聚类性能保持稳定。因此，建议将 λ_1 和 λ_2 的范围从 0.1 微调到 1。

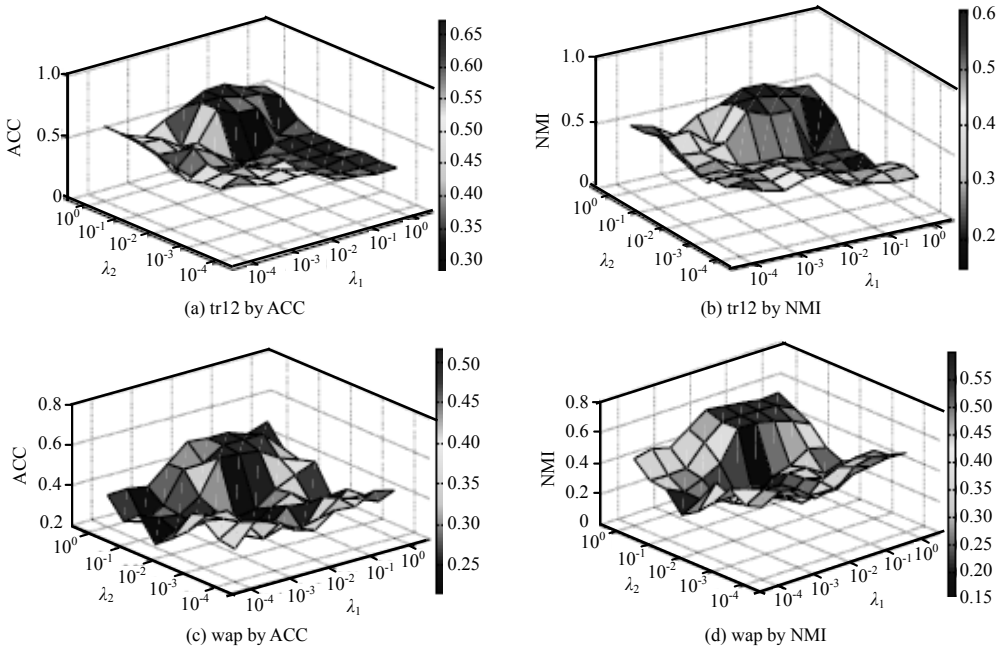


(a) convergence

(b) clustering performance by ACC

(c) clustering performance by NMI

Fig.1 ISEC algorithm performance
图1 ISEC算法性能



(a) tr12 by ACC

(b) tr12 by NMI

(c) wap by ACC

(d) wap by NMI

Fig.2 Parameter analysis on λ_1 and λ_2
图2 λ_1和λ_2对ISEC性能(ACC和NMI)的影响

4 结论

聚类是一种主要的数据挖掘手段，针对现有聚类方法的不足，提出了一种改进的谱组合聚类(ISEC)方法。该方法将聚类问题建模为基于共协矩阵的图分割问题，然后通过采用增强拉格朗日乘数法来进行优化求解。在 12 个真实数据集上获得的实验结果验证了 ISEC 方法的有效性。下一步工作将研究面向缺失值处理的子空间聚类算法，进一步提升聚类的质量和拓展聚类的应用价值。

参考文献：

[1] 陈强. 基于聚类技术的多阈值图像分割技术[J]. 太赫兹科学与电子信息学报, 2018,16(4):715-718. (CHEN Qiang. Multi-threshold image segmentation based on clustering method[J]. Journal of Terahertz Science and Electronic Information Technology, 2018,16(4):715-718.)

[2] PERROTTA C,WILLIAMSON B. The social life of learning analytics:cluster analysis and the 'performance' of algorithmic education[J]. Learning,Media and Technology, 2018,43(1):3-16.

[3] HUANG D,LAI J H,WANG C D. Robust ensemble clustering using probability trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2016,28(5):1312-1326.

[4] SHEN J,HAO X,LIANG Z,et al. Real-time superpixel segmentation by DBSCAN clustering algorithm[J]. IEEE Transactions on Image Processing, 2016,25(12):5933-5942.

[5] YI J,YANG T,JIN R,et al. Robust ensemble clustering by matrix completion[C]// 2012 IEEE 12th International Conference on Data Mining. [S.l.]:IEEE, 2012:1176-1181.

[6] LI E,LI Q,GENG Y,et al. Ensemble clustering using maximum relative density path[C]// 2018 IEEE International Conference on Big Data and Smart Computing(BigComp). [S.l.]:IEEE, 2018:190-197.

[7] SUN Z,SONG Q,ZHU X,et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5):1623-1637.

[8] LIU H,SHAO M,LI S,et al. Infinite ensemble for image clustering[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]:ACM, 2016:1745-1754.

[9] TOPCHY A,JAIN A K,PUNCH W. Clustering ensembles:models of consensus and weak partitions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27(12):1866-1881.

[10] WU J,LIU H,XIONG H,et al. K-means-based consensus clustering:a unified view[J]. IEEE Transactions on Knowledge and Data Engineering, 2015,27(1):155-169.

[11] SHERI A M,RAFIQUE M A,HASSAN M T,et al. Boosting discrimination information based document clustering using consensus and classification[J]. IEEE Access, 2019(7):78954-78962.

[12] VEGA-PONS S,CORREA-MORRIS J,RUIZ-SHULCLOPER J. Weighted partition consensus via kernels[J]. Pattern Recognition, 2018,43(8):2712-2724.

[13] LIU H,WU J,LIU T,et al. Spectral ensemble clustering via weighted K-means:theoretical and practical evidence[J]. IEEE Transactions on Knowledge and Data Engineering, 2017,29(5):1129-1143.

[14] ZHOU P,DU L,WANG H,et al. Learning a robust consensus matrix for clustering ensemble via Kullback-Leibler divergence minimization[C]// 24th International Joint Conference on Artificial Intelligence. [S.l.]:IEEE, 2015:4112-4118.

[15] WANG Y,WU L,LIN X,et al. Multiview spectral clustering via structured low-rank matrix factorization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018,29(10):4833-4843.

[16] LU C,FENG J,LIU W,et al. Tensor robust principal component analysis with a new tensor nuclear norm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,15(2):289-302.

[17] KANZOW C,STECK D,WACHSMUTH D. An augmented Lagrangian method for optimization problems in Banach spaces[J]. SIAM Journal on Control and Optimization, 2018,56(1):272-291.

[18] LEPENDU M,JIANG X,GUILLEMOT C. Light field inpainting propagation via low rank matrix completion[J]. IEEE Transactions on Image Processing, 2018,27(4):1981-1993.

[19] FANG X,XU Y,LI X,et al. Robust semi-supervised subspace clustering via non-negative low-rank representation[J]. IEEE Transactions on Cybernetics, 2015,46(8):1828-1838.

[20] JEYASUDHA A,PRIYA K. Object recognition based on LBP and discrete wavelet transform[J]. International Journal of Advances in Signal and Image Sciences, 2016,2(1):24-30.

[21] LIU H,LIU T,WU J,et al. Spectral ensemble clustering[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]:ACM, 2015:715-724.

[22] YAN D,HUANG L,JORDAN M I. Fast approximate spectral clustering[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]:ACM, 2009:907-916.

[23] SHAO M,LI S,DING Z,et al. Deep linear coding for fast graph clustering[C]// 24th International Joint Conference on Artificial Intelligence. [S.l.]:ACM, 2015:3798-3803.