

1. 引言

1.1 选题背景

现有的GraphSC及其改进方法中的拉普拉斯图都是预先定义且固定不变的，且预先定义的拉普拉斯图往往由KNN算法直接生成。然而，KNN方法构造的图有可能会破坏部分原始数据的局部连通性，从而使得相关算法的性能不能达到最优。针对这个问题，我们对GraphSC算法进行了改进，提出了自适应正则化稀疏编码（Graph Regularization Sparse Coding with Adaptive Neighbour, GraphSCAN）算法。

1.2 相关技术

GraphSC将拉普拉斯图引入SC算法，并使用图拉普拉斯正则项来保留数据局部流形结构。

对于给定一组 n 维数据点 $\mathbf{x}_1, \dots, \mathbf{x}_m$ ，构造一个有 m 个顶点的最近邻图 G ，其中每个顶点代表一个数据点。令 W 为图的权重，若 $\mathbf{x}_i$ 是 $\mathbf{x}_j$ 的最近邻或者 $\mathbf{x}_j$ 是 $\mathbf{x}_i$ 的最近邻，那么 $W_{ij}=1$ ，否则 $W_{ij}=0$ 。定义 $\mathbf{x}_i$ 的度为 $d_i = \sum_{j=1}^m W_{ij}$ ，且 $D = \text{diag}(d_1, \dots, d_m)$ 。

图正则化目标函数可表示为：

$$\frac{1}{2} \sum_{i,j=1}^m (\mathbf{e}_i - \mathbf{e}_j)^T W_{ij} = \text{Tr}(\mathbf{E}^T \mathbf{L} \mathbf{E})$$

其中 $\mathbf{L}$ 是图拉普拉斯矩阵， $\mathbf{L} = \mathbf{D} - \mathbf{S}$ ，GraphSC 的目标函数为：

$$\begin{cases} \min \|\mathbf{X} - \mathbf{A}\mathbf{E}\|_F^2 + \beta \sum_{i=1}^m \|\mathbf{e}_i\|_1 + \lambda \text{tr}(\mathbf{E}^T \mathbf{L}_S \mathbf{E}) \\ \text{s.t. } \lambda \geq 0, \|\mathbf{a}_i\|^2 \leq c, i = 1, \dots, k \end{cases}$$

2. 自适应图正则化稀疏编码目标函数

设输入的原始数据集为 $\mathbf{X}$ ， $\mathbf{X}$ 中的任意一数据点 $\mathbf{x}_i$ 都有所有数据点 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ 可以作为近邻与 $\mathbf{x}_i$ 连接，连接的概率为 s_i 。基于数据点间的欧几里得距离构造邻域矩阵，距离越小则成为最近邻的可能性越大。在此给出本文算法的求解形式如下：

$$\begin{aligned} \min & \|\mathbf{X} - \mathbf{A}\mathbf{E}\|_F^2 + \alpha \left(\sum_{j=1}^n \left(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 s_{ij} + \gamma s_{ij}^2 \right) \right) + \beta \sum_{i=1}^n \|\mathbf{e}_i\|_1 + \lambda \text{tr}(\mathbf{E}^T \mathbf{L}_S \mathbf{E}) \\ \text{s.t. } & \lambda \geq 0, \|\mathbf{a}_i\|^2 \leq c, s_i^T \mathbf{1} = 1, 0 \leq s_i \leq 1 \end{aligned}$$

3. 自适应图正则化稀疏编码算法

算法2. 自适应图稀疏编码算法

输入：数据矩阵 $\mathbf{X} \in \mathbb{R}_+^{m \times n}$ 的初始值，簇号 c ，参数 $k, \lambda, \alpha, \beta$ ；

输出：字典矩阵 $\mathbf{A}$ 、稀疏编码 $\mathbf{E}$ 。

Step1 初始化字典矩阵 $\mathbf{A}$ 、稀疏编码 $\mathbf{E}$ ；

Step2 重复

通过稀疏编码算法1更新字典矩阵 $\mathbf{A}$ 、稀疏编码 $\mathbf{E}$ ；

3) 对于每个 i ，通过求解式 $\mathbf{R}(s_i, \eta, \zeta_i) = \frac{1}{2} \left\| s_i + \frac{1}{2\gamma_i} \mathbf{d}_i^x \right\|_2^2 - \eta (s_i^T \mathbf{1} - 1) - \zeta_i^T s_i$ 更新 $\mathbf{S}$ 的第 i 行。

直到收敛；

Step3 求得结果。

4. 实验

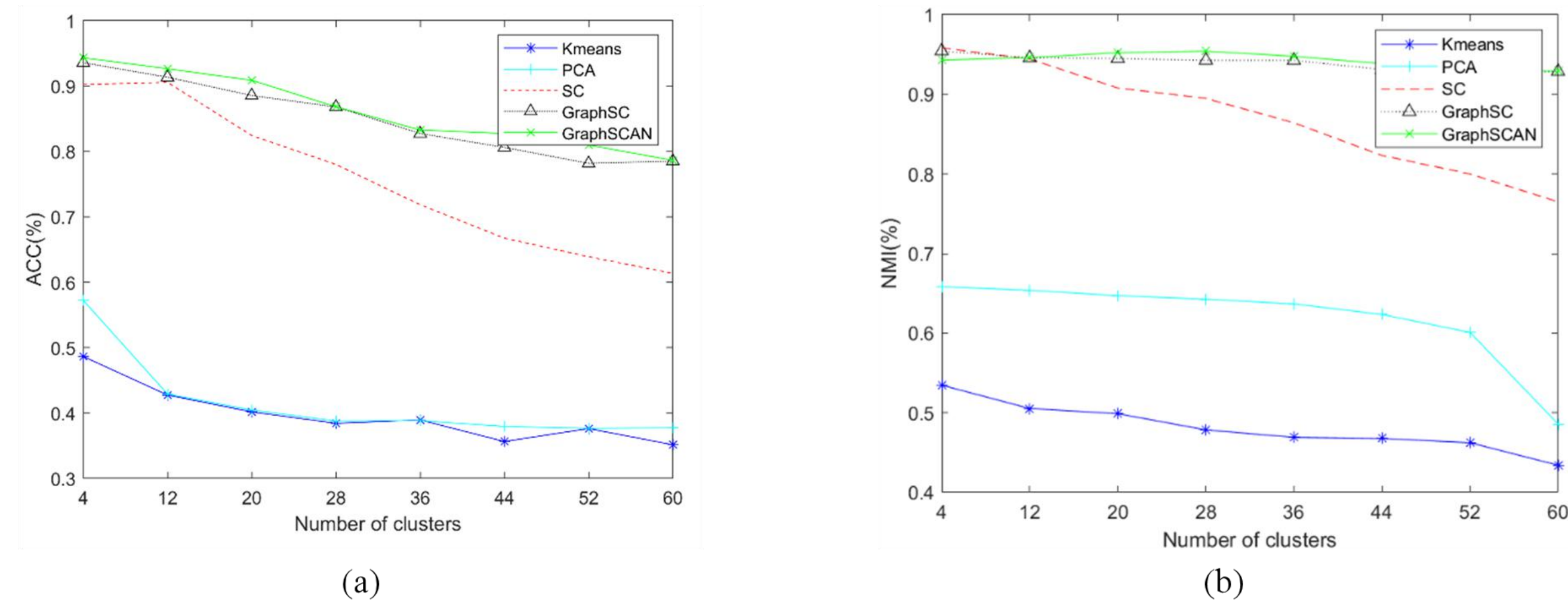


图 1 (a) ACC (b) NMI 在 CMU-PIE 数据集上的簇数关系图

Fig.1 (a) Accuracy. (b) Normalized mutual information versus. the number of clusters on CMU-PIE data set

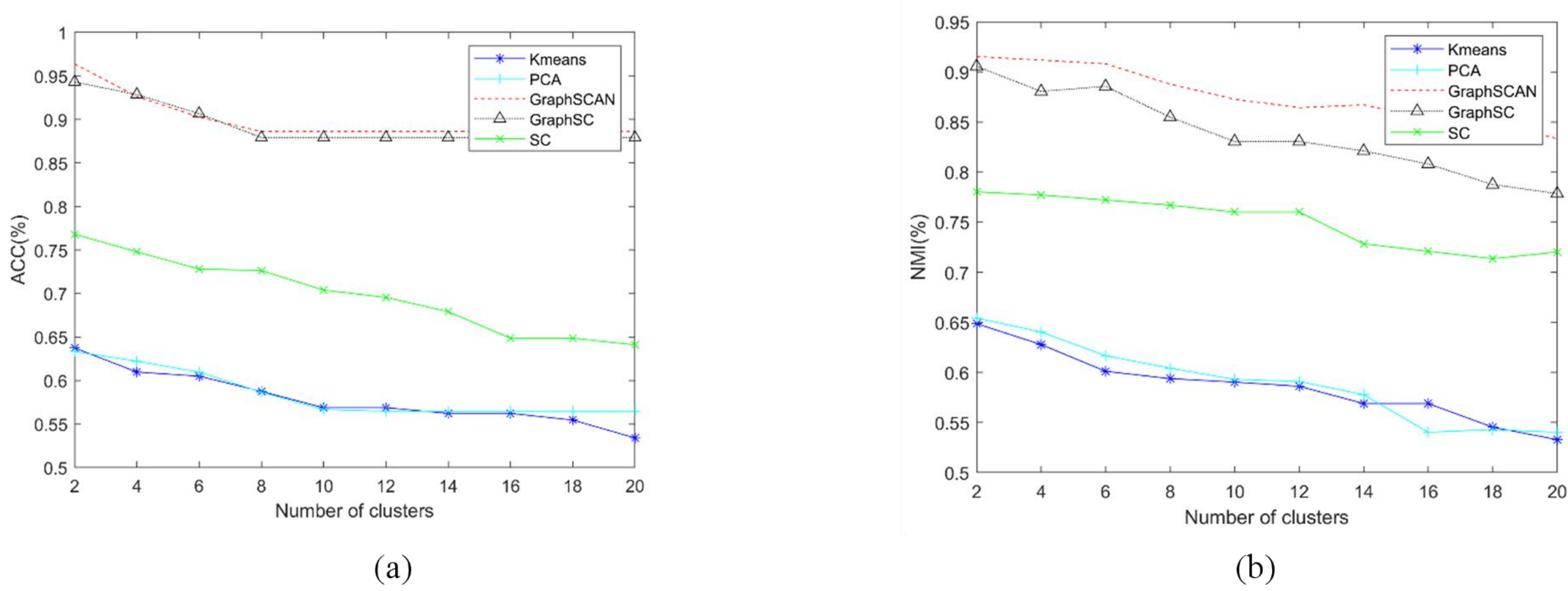


图 2 (a) ACC (b) NMI 在 COIL20 数据集上的簇数关系图

Fig.2 (a) Accuracy. (b) Normalized mutual information versus. the number of clusters on COIL20 data set

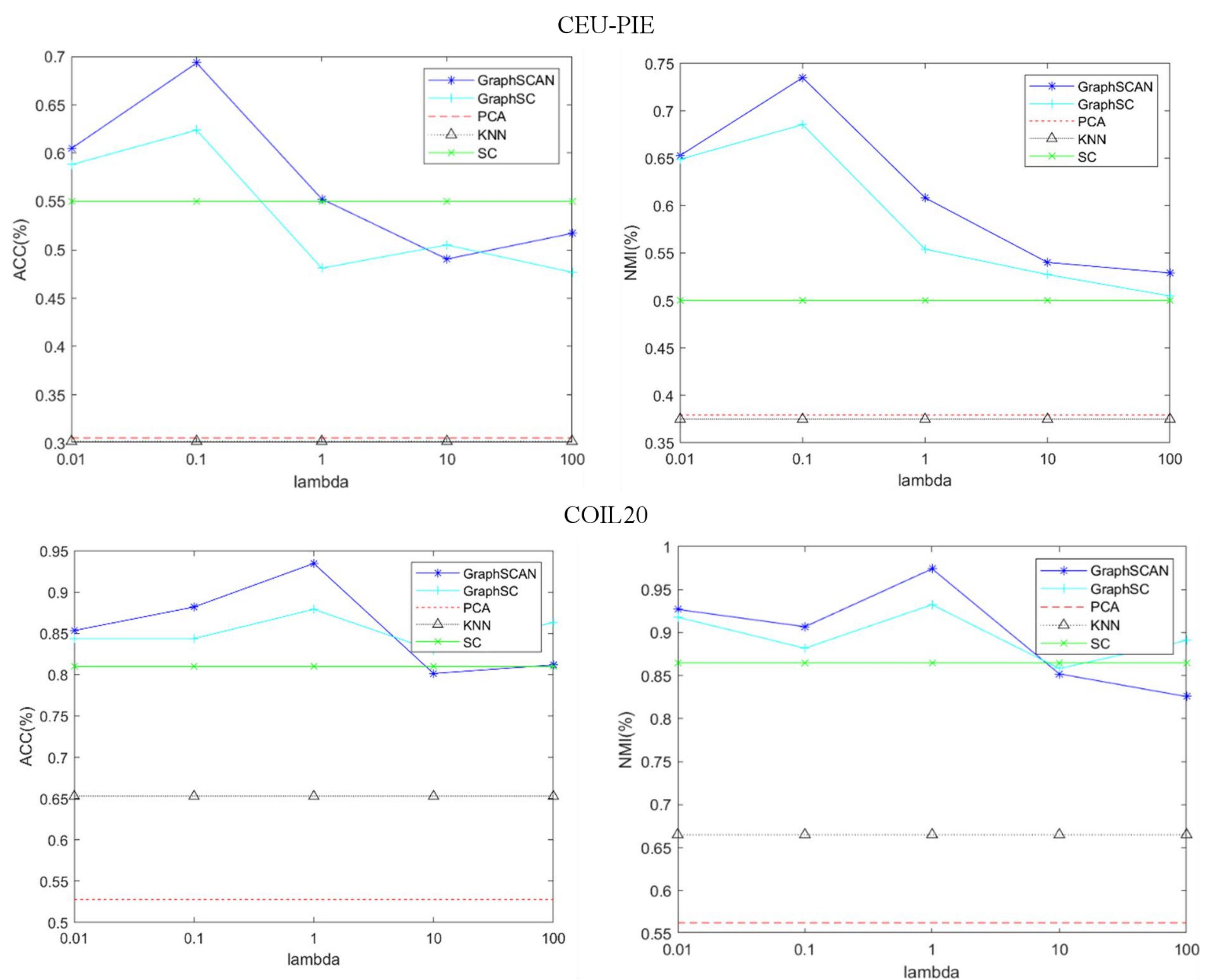


Fig. 3 The influence of λ parameters on different algorithms

图 3 各算法中 λ 参数对性能的影响