CCML 2021 第十八届中国机器学习会议
The 18th China Conference on Machine Learning
2021年8月6-8日 中国·长沙
AUG 6-8, 2021 CHANGSHA, CHINA

自适应图正则的单步子空间聚类

引言：

传统的子空间聚类遵循两步的方法，首先从数据矩阵中学习相似矩阵然后再进行谱聚类， 聚类结果高度依赖于数据的相似性学习。 由于相似性度量和数据聚类通常分为两个步骤进行，仍然存在一些缺陷，如：没有利用相似性学习和聚类相互依赖的事实，可能导致较少的全局最优解；特征分解的过程非常耗时，未考虑局部光滑性等。因此提出了一个基于局部图正则的联合优化框架--自适应图正则的单步子空间聚类。

方法：

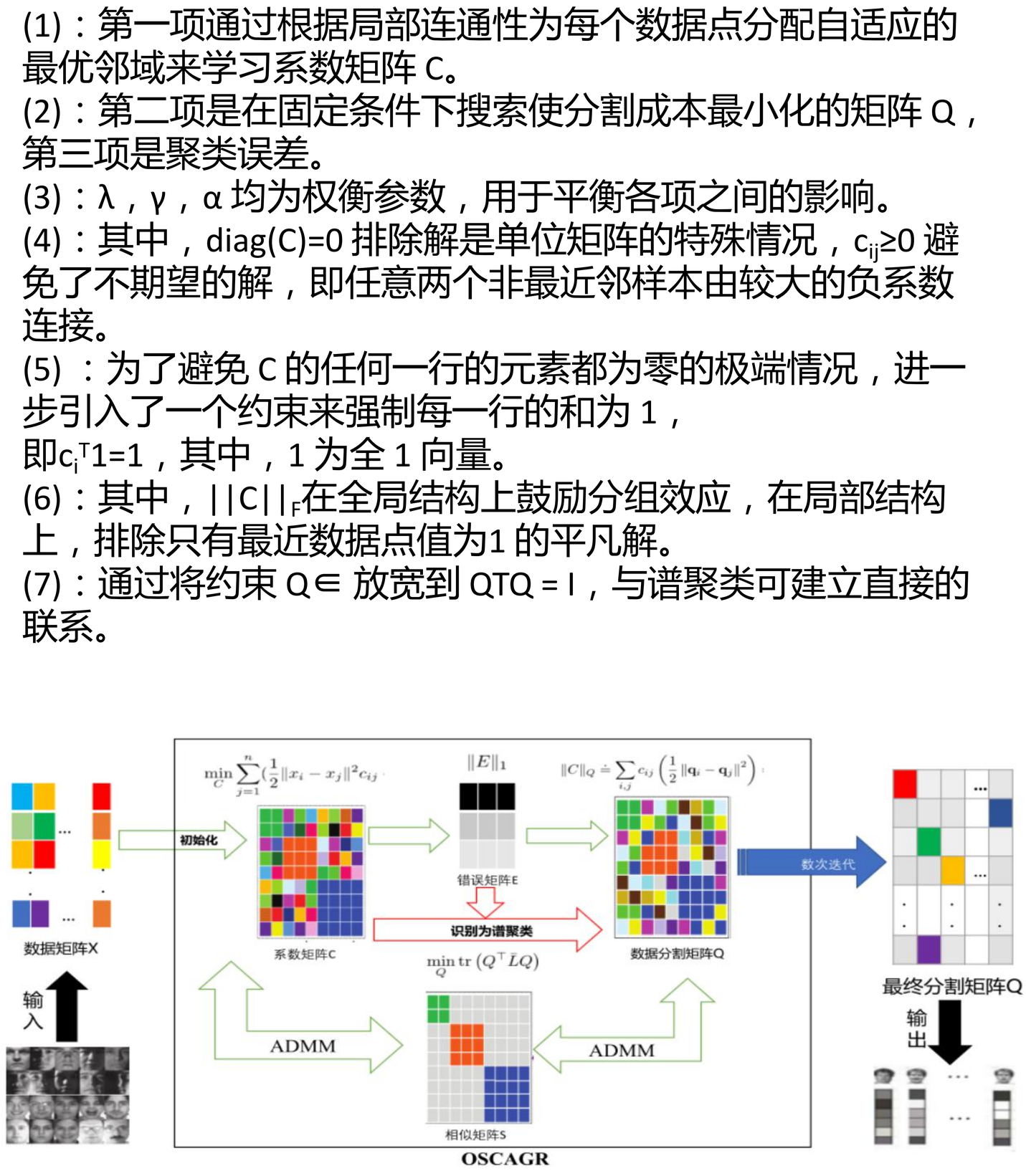
自适应图正则　量化范数约束　误差

$$\min_{C,E,Q}\sum_{ij}\| x_i - x_j \|^2 c_{ij} + \gamma \| C \|_F^2 + \alpha \| q_i - q_j \|^2 c_{ij} + \lambda \| E \|_1$$

单步子空间聚类

$$\text{s.t.} X = XC + E,\quad c_i^T \mathbf{1} = \mathbf{1},\quad c_{ij} \geq 0,\quad diag(C) = 0,\quad Q \in \mathcal{Q}$$

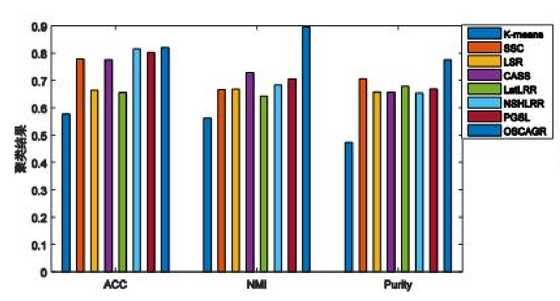
(1)：第一项通过根据局部连通性为每个数据点分配自适应的最优邻域来学习系数矩阵 C。
(2)：第二项是在固定条件下搜索使分割成本最小化的矩阵 Q，第三项是聚类误差。
(3)：λ，γ，α 均为权衡参数，用于平衡各项之间的影响。
(4)：其中，diag(C)=0 排除解是单位矩阵的特殊情况，c_{ij}≥0 避免了不期望的解，即任意两个非最近邻样本由较大的负系数连接。
(5) ：为了避免 C 的任何一行的元素都为零的极端情况，进一步引入了一个约束来强制每一行的和为 1，
即c_i^T1=1，其中，1 为全 1 向量。
(6)：其中，||C||$_F$在全局结构上鼓励分组效应，在局部结构上，排除只有最近数据点值为1 的平凡解。
(7)：通过将约束 Q∈ 放宽到 QTQ = I，与谱聚类可建立直接的联系。



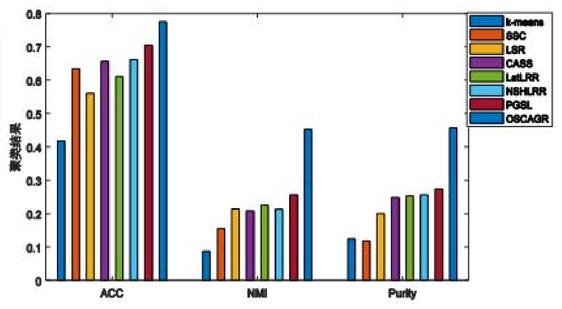
结论：

本文利用 Frobenius 范数鼓励分组效应，并根据局部连通性为每个数据点分配自适应的最优邻域来学习系数矩阵。同时考虑全局结构和局部结构，基于自表达模型， 保证数据空间中相近的点拥有较大的表示系数， 通过量化范数将子空间聚类两个独立的阶段整合到一个统一的优化框架中。最后，在多种类型数据集上进行大量实验，证明了所提算法的优越性。

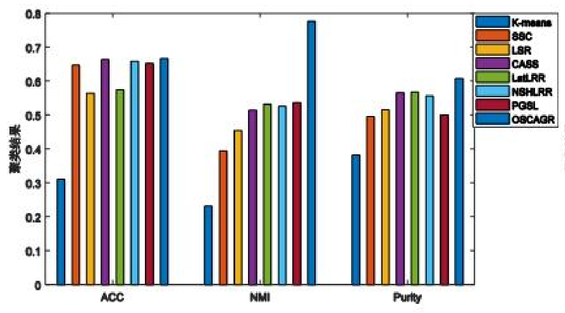
实验结果：



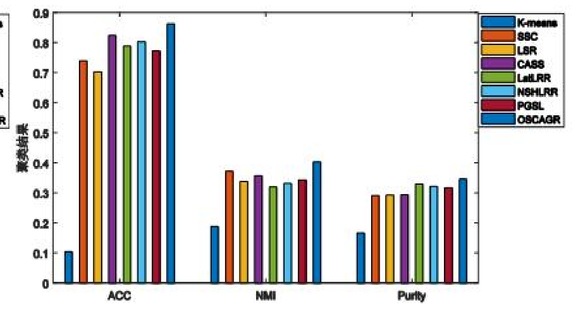
(a) COIL20



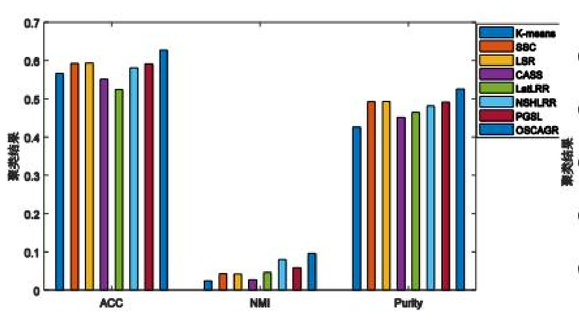
(b) MNIST



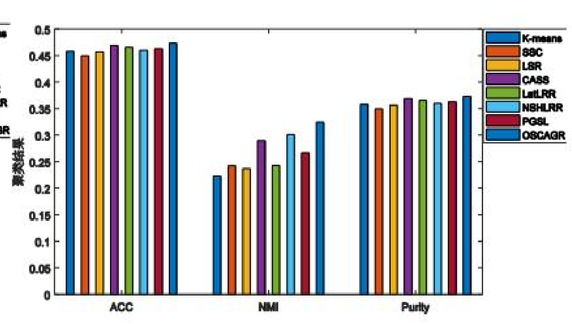
(c) ORL



(d) YaleB



(e) Isolet



(f) Vehicle

不同数据集中的准确度（%）

Data	k-means	SSC	LSR	CASS	LatLRR	NSHLRR	PGSL	OSCAGR
COIL20	57.65	77.82	66.34	75.11	65.57	81.48	80.31	**81.96**
MNIST	41.78	63.44	56.04	65.65	61.08	66.15	70.47	**77.51**
ORL	31.05	64.73	56.37	66.31	57.44	65.75	65.15	**66.64**
YaleB	10.36	73.89	70.22	82.41	78.88	80.24	77.20	**86.24**
Isolet	56.56	59.24	59.30	55.13	52.44	58.11	59.10	**62.68**
Vehicle	45.77	44.96	45.65	46.90	46.55	45.98	46.26	**47.32**

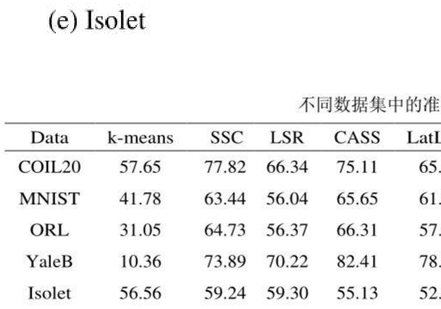
不同数据集中的标准化互信息（%）

Data	k-means	SSC	LSR	CASS	LatLRR	NSHLRR	PGSL	OSCAGR
COIL20	56.19	66.61	66.83	72.90	64.25	68.35	70.53	**89.56**
MNIST	8.79	15.54	21.41	20.88	22.56	21.31	25.59	**45.31**
ORL	23.09	39.41	45.38	51.40	53.17	52.59	53.62	**77.58**
YaleB	18.79	37.24	33.81	35.63	31.90	33.11	34.26	**40.29**
Isolet	2.39	4.32	4.24	2.69	4.58	7.96	5.85	**9.64**
Vehicle	22.24	24.24	23.69	28.96	24.26	30.10	26.65	**32.45**

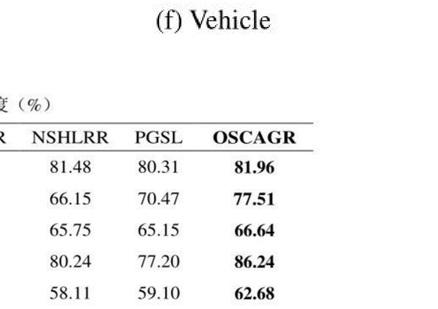
不同数据集中的纯度（%）

Data	k-means	SSC	LSR	CASS	LatLRR	NSHLRR	PGSL	OSCAGR
COIL20	47.28	70.60	65.72	65.64	67.85	65.38	66.84	**77.50**
MNIST	12.46	11.88	20.01	24.84	25.27	25.62	27.45	**45.77**
ORL	38.11	49.57	51.54	56.52	56.69	55.59	49.96	**60.75**
YaleB	16.70	29.06	29.21	29.32	32.90	32.11	31.65	**34.54**
Isolet	42.56	49.24	49.30	45.13	46.44	48.11	49.10	**52.56**
Vehicle	35.77	34.96	35.65	36.90	36.55	35.98	36.26	**37.32**

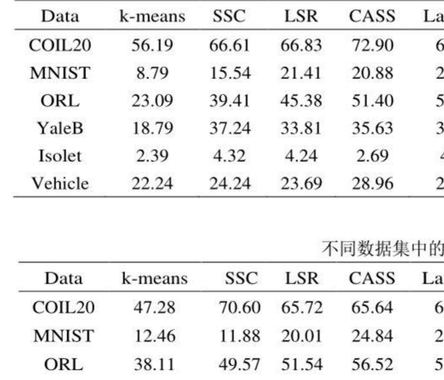
参数分析：



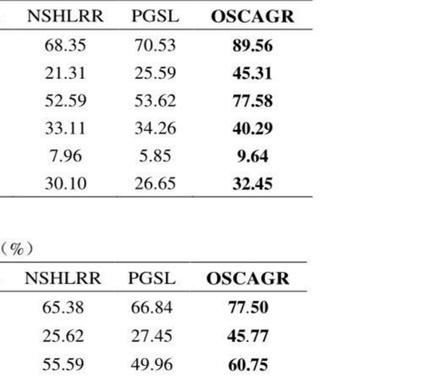
(a) COIL20



(b) YaleB



(c) COIL20



(d) YaleB

投稿人　姓名：程业超，刘惊雷
单位：烟台大学计算机与控制工程学院