



研究与开发

抽样子空间约束改进大数据谱聚类算法

聂茹

(华南理工大学广州学院电子信息工程学院, 广东 广州 510800)

摘要: 在分析经典谱聚目标函数与加权核 k -means 目标函数等价基础上, 设计了一种基于抽样子空间约束的改进大规模数据谱聚类算法, 算法通过加权核 k -means 迭代优化避免矩阵特征分解的大量资源被占用, 通过数据抽样及聚类中心的子空间约束, 避免全部核矩阵都被使用, 从而降低经典算法的时间空间复杂度。理论分析和实验结果表明, 改进算法保持与经典算法相近聚类精度, 提高了聚类效率, 验证了改进算法的有效性。

关键词: 大规模数据谱聚类; 加权核 k -means 算法; 数据抽样; 矩阵特征分解; 核矩阵

中图分类号: TP181

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018277

Improved large data spectral clustering algorithm based on sampling subspace constraint

NIE Ru

Electronic Information Engineering Institute, Guangzhou College of South China University of Technology, Guangzhou 510800, China

Abstract: On the basis of analyzing the equivalent function of the objective function of classical spectral clustering algorithm and the weighted kernel k -means objective function, an improved large-scale data spectrum clustering algorithm based on sampling subspace constraint was designed, the weighted kernel k -means iterative optimization was used to avoid the large resource consumption of Laplacian matrix feature decomposition, and by using data sampling and constraining the cluster center to the subspace generated by the sampling points, the use of all kernel matrices was avoided, thereby reducing the time-space complexity of classical algorithms. Theoretical analysis and experimental results show that the improved algorithm can greatly improve the clustering efficiency on the basis of maintaining similar clustering accuracy with the classic algorithm and verify the effectiveness of the proposed algorithm.

Key words: large scale data spectral clustering, weighted kernel k -means algorithm, data sampling, matrix feature decomposition, kernel matrix

1 引言

信息科技的快速发展产生 TB 甚至 PB 级的海

量信息数据, 如何提取有价值的知识显得尤为重要, 聚类是数据挖掘的主要工具, 能够将相似特征数据进行归类, 从而高效组织数据^[1]。现有的 K

收稿日期: 2018-06-13; 修回日期: 2018-10-31

基金项目: 广东省教育厅青年创新人才基金资助项目 (No.2016KQNCX227)

Foundation Item: Youth Innovation Talent Project of Guangdong Provincial Department of Education (No.2016KQNCX227)



均值、EM 算法及其改进等大多基于欧氏距离,当样本集非凸时,易陷入局部最优,基于非线性核距离计算的谱聚类,能实现包含非凸结构等任意形状样本集的全局最优解,比传统 k -means 等方法更有效,成为近年来极具竞争力的聚类方法^[2]。

但随着数据集规模的增大,经典谱聚类相似矩阵的计算和存储资源占用及特征分解的时间复杂度急剧增大^[3]。为此, Yan 等人^[4]采用快速逼近思想,先对代表数据进行聚类,然后根据代表点与普通数据间的一致性实现聚类;钱鹏江等人^[5]基于自适应性 GRC 谱聚类算法,通过约束谱聚类的指示向量并融合 CCMEB 理论将算法在大数据集上的复杂度降低到渐近线性时间复杂度。杨艺等人^[6]通过在选取的核心点集上分组和谱聚类,根据一致性实现大数据谱聚类。Kumar 等人^[7]根据 Nyström 误差界分析,以牺牲相似性计算准确性换来聚类的快速计算,提高了算法的效率。Song 等人^[8-9]比较了近似相似度矩阵计算的 Nyström 方法和稀疏化方法,并通过保留稀疏矩阵及并行化策略,实现在大型数据集上的谱聚类。

为避开 Laplacian 矩阵特征分解的时间复杂度, Dhillon 等人^[10]在研究加权核 k -means 算法与谱聚类目标函数等价的基础上,通过加权核 k -means 迭代优化目标函数,在大数据下矩阵特征分解困难情况下,取得了更好的聚类结果。但标准核距离在有效捕捉非线性数据结构的同时,还需存储一个较大的核矩阵,限制了其在大规模数据集中的应用。为此在剖析核矩阵复杂性原因的基础上,提出基于抽样子空间约束的改进大数据谱聚类方法,算法通过约束样本点生成的子空间来逼近聚类中心,从而仅需计算部分核矩阵,极大减少核矩阵的复杂度,理论分析和实验结果验证了算法的有效性。

2 目标函数与加权核 k -means 等价性分析

经典谱聚类基于谱图理论,将数据作为无

向加权图,从而将聚类问题转变为最佳图划分问题^[4]。

2.1 基于谱图理论的经典谱聚类

将数据集表示为无向加权图 $G=(V, E, A)$, 其中, V 表示数据集, E 表示任意两个数据节点之间的连接边集合,而每个连接边都赋予一个衡量两节点同类可能性大小的权重,矩阵 A 则为这些连接边权重值的集合,其非负且对称。

将 n 点数据集 $X=\{x_1, x_2, \dots, x_n\}$ 视为无向图,则将其聚类为 k 类,即在 E 和 A 约束下将 $V=X$ 划分为 k 个子集,即 $V=\bigcup_{i=1}^k V_i$ 且 $V_i \cap V_j = \emptyset, i \neq j$ 。

高性能谱聚类应使其划分后的类内连接权值最大或类间连接权值最小,由此得到基于归一化图划分的谱聚类的目标函数^[8]:

$$\max J(V) = \text{NAssos}(V_1, \dots, V_k) = \sum_{i=1}^k \text{Nlinks}(V_i, V_i) \quad (1)$$

其中, $\text{Nlinks}(V_i, V_j)$ 描述了任意两个子类 $V_i, V_j \subset V$ 间元素连接边的归一化总权值和,即:

$$\text{Nlinks}(V_1, V_2) = \frac{\text{links}(V_1, V_2)}{\text{links}(V_1, V)} = \frac{\sum_{i \in V_1, j \in V_2} A_{i,j}}{\sum_{i \in V_1, l \in V_2} A_{i,l}} \quad (2)$$

建立类别隶属矩阵 $U=\{u_1, u_2, \dots, u_k\}^T \in \mathbb{R}^{k \times n}$, u_i 表示元素隶属,即若 $x_j \in V_i$, 则 $U_{ij}=1$, 否则 $U_{ij}=0$, 根据节点类别的唯一性有 $U^T 1_k = 1_n$, 向量 $1_n \in \mathbb{R}^{n \times 1}$ 的元素全为 1, 则式 (1) 中经典谱聚类的目标函数可表示为:

$$\begin{aligned} \text{NAssoc}(V_1, \dots, V_k) &= \sum_{i=1}^k \text{Nlinks}(V_i, V_i) = \sum_{i=1}^k \frac{u_i^T A u_i}{u_i^T D u_i} \\ &= \sum_{i=1}^k \left(\frac{u_i^T}{(u_i^T D u_i)^{1/2}} A \frac{u_i}{(u_i^T D u_i)^{1/2}} \right) \\ &= \sum_{i=1}^k z_i^T A z_i = \text{tr}(\tilde{Z} D^{-1/2} A D^{-1/2} \tilde{Z}^T) \end{aligned} \quad (3)$$

其中, 类间连接重写为 $\text{links}(V_i, V_j) = u_i^T A u_j$, $\text{links}(V_i, V) = u_i^T D u_i$, $D = \text{diag}(A 1_n)$, $\text{diag}(\cdot)$ 表示对角矩阵, $Z = (U D U^T)^{-1/2} U \in \mathbb{R}^{k \times n}$, $\tilde{Z} = Z D^{1/2}$ 且

满足约束 $\tilde{Z}\tilde{Z}^T = I_k$, $I_k \in \mathbb{R}^{k \times k}$ 是单位矩阵, $\text{tr}(\cdot)$ 表示矩阵迹。式 (3) 表明最大化谱聚类目标函数 $\text{NAssoc}(V_1, \dots, V_2)$, 等价于最大化矩阵的迹 $\max_Z (\text{tr}(\tilde{Z}D^{-1/2}AD^{-1/2}\tilde{Z}^T))^{[10]}$ 。

2.2 加权核 k -means 算法

加权核 k -means 算法是泛化的 k -means 算法^[11], 其聚类目标函数为使分类中数据点到类中心的距离误差加权总和最小, 即:

$$\begin{aligned} \min J(V) &= \sum_{i=1}^k \sum_{x_j \in V_i} w_j |\kappa(x_j, \cdot) - c_i(\cdot)|_{H_k}^2 \\ &= \sum_{i=1}^k \sum_{j=1}^k w_j U_{ij} |\kappa(x_j, \cdot) - c_i(\cdot)|_{H_k}^2 \end{aligned} \quad (4)$$

其中, $\kappa(\cdot): \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ 表示核距离函数, H_k 表示由 $\kappa(\cdot)$ 生成的再生核希尔伯特空间, $|\cdot|_{H_k}^2$ 为其范数, $c_i(\cdot)$ 为类中心。 w_i 为对每个数据 x_i 分配的权值, 由权值矩阵 $W = \text{diag}(w_1, \dots, w_n)$ 可构建加权类别矩阵 $Y = (y_1, \dots, y_k)^T = UW$ 及其归一化矩阵 $\hat{Y}$ 和 $\tilde{Y}$:

$$\begin{cases} \hat{Y} = [\text{diag}(s_1, \dots, s_k)]^{-1} UW = LUW = LY \\ \tilde{Y} = [\text{diag}(\sqrt{s_1}, \dots, \sqrt{s_k})]^{-1} UW^{1/2} = L^{1/2} UW^{1/2} \end{cases} \quad (5)$$

其中, $L = [\text{diag}(s_1, \dots, s_k)]^{-1}$, $s_i = y_i^T \mathbf{1}_n$ 。则加权核 k -means 最佳的类中心可以表示为^[11]:

$$c_i(\cdot) = \frac{1}{s_i} \sum_{x_j \in V_i} w_j \kappa(x_j, \cdot) = \sum_{j=1}^n \hat{Y}_{ij} \kappa(x_j, \cdot) \quad (6)$$

与式 (3) 相似, 将式 (4) 目标函数平方项展开得到目标函数的矩阵迹形式为:

$$\begin{aligned} \min J(V) &= \text{tr}(WK) + \sum_{i=1}^k (s_i \hat{Y}_i^T K \hat{Y}_i - 2y_i^T K \hat{Y}_i) \\ &= \text{tr}(W^{1/2} K W^{1/2}) - \text{tr}(\tilde{Y} W^{1/2} K W^{1/2} \tilde{Y}) \end{aligned} \quad (7)$$

其中, $K \in \mathbb{R}^{n \times n}$ 是核函数矩阵, $K_{ij} = \kappa(x_i, x_j)$, 由于式 (7) 中前一项 $\text{tr}(W^{1/2} K W^{1/2})$ 为常量, 因此得到式 (7) 中加权核 k -means 的目标函数 $J(V)$ 等价于 $\max \text{tr}(\tilde{Y} W^{1/2} K W^{1/2} \tilde{Y}^T)$, 且有:

$$\begin{aligned} \text{tr}(\tilde{Y} W^{1/2} K W^{1/2} \tilde{Y}^T) &= \text{tr}(\tilde{Z} D^{1/2} \cdot D^{-1} A D^{-1} \cdot D^{1/2} \tilde{Z}^T) \\ &= \text{tr}(\tilde{Z} D^{-1/2} A D^{-1/2} \tilde{Z}^T) \end{aligned} \quad (8)$$

根据上述推导可见, 基于矩阵迹, 经典谱聚类的 Normalized Cut 目标函数与加权核 k -means 目标函数等价, 因此可以通过加权核 k -means 的迭代更新快速求解最佳的子图划分, 避免经典谱聚类矩阵特征分解的巨大资源占用。

3 改进加权核 k -means 算法

加权核 k -means 算法需要计算和存储 $n \times n$ 的核矩阵 K , 在大规模数据集应用时仍受到限制。为此, 文中通过随机抽样近似计算核矩阵并约束样本子空间来逼近聚类中心, 设计了改进加权核 k -means 算法, 以减少大数据集应用中时间和空间复杂度。

设由随机采样 $\hat{X} = \{\hat{x}_1, \dots, \hat{x}_m\}$ 构造的子空间为 $\hat{H}_k = \text{span}(\kappa(\hat{x}_1, \cdot), \dots)$, $m \ll n$, 则在 $\hat{H}_k$ 下的加权核 k -means 算法目标函数为:

$$\begin{aligned} \min J(V) &= \\ \sum_{i=1}^k \sum_{j=1}^n w_j U_{ij} |\kappa(x_j, \cdot) - c_i(\cdot)|_{H_k}^2, c_i(\cdot) \in \hat{H}_k \end{aligned} \quad (9)$$

根据式 (6) 构造采样子空间聚类中心为 $c_i(\cdot) = \sum_{j=1}^m \alpha_{ij} \kappa(\hat{x}_j, \cdot)$, $\alpha_{ij} \in \mathbb{R}^{k \times m}$ 为未知参数, 则式 (9) 可展开为:

$$\begin{aligned} \min J(V) &= \\ \text{tr}(WK) + \sum_{i=1}^k (s_i \alpha_i^T \hat{K} \alpha_i - 2y_i^T \tilde{K} \alpha_i) \end{aligned} \quad (10)$$

其中, $\text{tr}(WK)$ 为常数, 当式 (17) 取最小值时, $\beta_i = s_i \alpha_i^T \hat{K} \alpha_i - 2y_i^T \tilde{K} \alpha_i$ 对 α_i 的偏导数为 0, 根据矩阵求偏导公式可得最优解 $\alpha_i = \hat{K}^{-1} \tilde{K}^T \hat{Y}_i$, 即:

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)^T = \hat{Y} \tilde{K} \hat{K}^{-1} \quad (11)$$

其中, $\hat{K} \in \mathbb{R}^{m \times m}$ 为 $\hat{X}$ 内样本间核相似矩阵, 而 $\tilde{K} \in \mathbb{R}^{n \times m}$ 为 $\hat{X}$ 内样本与 X 元素间核相似矩阵。由式 (11) 可得 $s_i \alpha_i^T \hat{K} \alpha_i = y_i^T \tilde{K} \alpha_i$, 代入式 (10) 并结合式 (5) 得到在子空间约束下加权核 k -means 目标函数为:



$$\begin{aligned}
J(V) &= \text{tr}(WK) + \sum_{i=1}^k (s_i \alpha_i^T \hat{K} \alpha_i - 2y_i^T \tilde{K} \alpha_i) \\
&= \text{tr}(WK) - \sum_{i=1}^k (y_i^T \tilde{K} \hat{K}^{-1} \tilde{K}^T \hat{y}_i) \\
&= \text{tr}(WK) - \text{tr}(\tilde{Y} \tilde{K} \hat{K}^{-1} \tilde{K}^T \hat{Y}^T) \\
&= \text{tr}(W^{1/2} K W^{1/2}) - \text{tr}(\tilde{Y} W^{1/2} \tilde{K} \hat{K}^{-1} \tilde{K}^T W^{1/2} \tilde{Y}^T) \quad (12)
\end{aligned}$$

根据式(12)进行大数据谱聚类时,其类别矩阵 U 可以通过 $\hat{K}$ 和 $\tilde{K}$ 计算,而 $\hat{K}$ 是 $\tilde{K}$ 的一部分,因此只需要计算 $\tilde{K}$ 即可以进行最终的聚类。由于采样点数 $m \ll n$, $\tilde{K}$ 的计算开销要远小于 K 。由抽样数据构建的 $\hat{K}$ 和 $\tilde{K}$ 计算数据集类别矩阵 U 的详细过程如下。

算法 1 抽样子空间约束改进谱聚类算法

输入: 大规则数据集 $X = \{x_1, \dots, x_n\}$; 随机采样数 m , $m \ll n$; 聚类数 k ; 最大迭代次数 $it_{\max}$;

输出: 包含最终聚类信息的类别矩阵 U 。

初始化:

随机选取样本集 $\hat{X} = \{\hat{x}_1, \dots, \hat{x}_m\}$;

计算 $\tilde{K} = [\kappa(x_i, \hat{x}_j)]_{n \times m}$ 和 $\hat{K} = [\kappa(\hat{x}_i, \hat{x}_j)]_{m \times m}$;

计算 $T = \tilde{K} \hat{K}^{-1}$;

矩阵 U 初始化为随机值; 初始迭代赋初值为 $it = 0$;

重复计算:

$it = it + 1$;

根据 $Y = UW$, 计算 $L = [\text{diag}(Y \mathbf{1}_n)]^{-1}$ 和 $\hat{Y} = LY$;

计算 $\alpha = \hat{Y} T = \hat{Y} \tilde{K} \hat{K}^{-1}$;

For $j = 1, \dots, n$ do

为每个点 x_j 找到最近的类中心:

$i^* = \arg \min_{i \in [k]} (\alpha_i^T \hat{K} \alpha_i - 2\varphi_j^T \alpha_i)$; //其中, α_i^T 是 α

的第 i 行, 而 φ_j^T 是 $\tilde{K}$ 的第 j 行。

更新 U 的第 j 列, 令 $i = i^*$ 行元素 $U_{ij} = 1$, 而其余元素为 0;

End For;

算法结束: 隶属矩阵 U 不再变化或 $it > it_{\max}$

4 改进算法性能分析

4.1 改进算法计算复杂度分析

抽样改进加权核 k -means 算法的计算复杂度主要集中在矩阵求逆 $\hat{K}^{-1}$ 和 $T = \tilde{K} \hat{K}^{-1}$, 时间复杂度为 $O(m^3 + m^2 n)$, 另外 α 和 U 的更新需要 $O(mnkt)$ 的时间开销 (t 为预设迭代次数), 改进算法的总复杂度为 $O(m^3 + m^2 n + mnkt)$ 。为避免矩阵求逆的影响, 可将 α 的计算转化成式(13):

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^{k \times m}} \left[\frac{1}{2} \text{tr}(\alpha \hat{K} \alpha^T) - \text{tr}(\hat{Y} \tilde{K} \alpha^T) \right] \quad (13)$$

式(13)采用梯度下降法的计算开销是 $O(m^2 k \log(1/\varepsilon))$ (ε 是期望的精度)。于是整个时间复杂度降低到 $O(m^2 k \log(1/\varepsilon) + m^2 n + mnkt)$, 且当 $k \log(1/\varepsilon) \ll m$ 时, $O(m^2 k \log(1/\varepsilon)) \ll O(m^3)$ 。改进算法中矩阵 $\tilde{K}$ 占用空间复杂度为 $O(mn)$, 比标准算法的 $O(n^2)$ ($m \ll n$) 大幅降低。

4.2 改进算法的误差范围分析

改进算法使用随机采样构建子空间 $\hat{H}_k$, 并将类中心约束在子空间中, 基于此分析其相对于原算法的聚类误差范围。用 $\xi = (\xi_1, \xi_2, \dots, \xi_n)^T \in \{0, 1\}^{n \times 1}$ 描述随机采样, 即当 x_i 被随机选取, 则 $\xi_i = 1$, 否则 $\xi_i = 0$, 则改进算法的聚类误差可表示为:

$$\begin{aligned}
\Phi(Y, \xi) &= \text{tr}(W^{1/2} K W^{1/2}) + \\
&\sum_{i=1}^k \left\{ \min_{\alpha_i \in \mathbb{R}^{n \times 1}} [s_i (\alpha_i \cdot \xi)^T K (\alpha_i \cdot \xi) - 2y_i^T K (\alpha_i \cdot \xi)] \right\} \quad (14)
\end{aligned}$$

根据前文相关变量的定义, 当 $W^{1/2} K W^{1/2}$ 为特殊特征值, 其非 0 特征值数多于数据集总类别数的 2 倍时, 可以推导出算法的误差满足:

$$\frac{E_{\xi}[\Phi(Y, \xi)] - \Phi(Y, \mathbf{1}_n)}{\Phi(Y, \mathbf{1}_n)} \leq \frac{2k}{m} \quad (15)$$

式(15)表明当 $W^{1/2} K W^{1/2}$ 的非 0 特征值数是总类别数 2 倍以上时, 基于抽样改进的加权核 k -means 算法的聚类误差会随着样本抽取点数的

增加以 $O(1/m)$ 的速率减小。

5 实验与分析

为验证文中提出算法（记为 Sik-SC）的性能，将其与经典谱聚类（Clas-SC）^[12-13]、基于标准加权核 k -means 谱聚类（Swk-SC）^[10]、基于 Nyström 近似谱聚类（Nys-SC）^[7] 和基于 MEKA 近似核聚类（Mek-SC）^[10] 4 种谱聚类方法进行性能比较实验。实验环境：Intel(R) Core i7-4710MQ @ 2.50 GHz，16 GB 内存，MATLAB 2016b。数据集的相关参数见表 1。

表 1 实验数据集相关参数

数据集	数据点个数	维数	类数
Waveform	5 000	40	3
Ringnorm	7 400	20	2
USPS	9 298	256	10
MNIST	70 000	784	10

Waveform 集^[14]由 3 类波形组成，其噪声均值为 0，方差为 1；Ringnorm 集^[15]由存在重叠的两种呈正态分布的样本组成，聚类难度较大；USPS 集和 MNIST 集为含有 10 种类型的手写数字图片集^[16]；实验中聚类结果的准确性采用归一化互信息（normalized mutual information, NMI）来评价：

$$NMI(U_c, U_t) = \frac{I(U_c, U_t)}{\sqrt{H(U_c) \cdot H(U_t)}} \quad (16)$$

其中， U_c 与 U_t 分别为实验得到类别矩阵和真实类别矩阵， $H(U_c)$ 、 $H(U_t)$ 是其信息熵，NMI 值位于 [0,1] 内，值越高说明聚类质量越好。实验聚类结果为 20 次实验结果的平均值。

5.1 聚类性能比较实验

如图 1 所示为 Nys-SC、Mek-SC 和 Sik-SC 3 种算法在 4 个数据集上不同采样点数时的实验结果，Clas-SC 与 Swk-SC 对所有数据进行处理，其在 4 个数据集上的 NMI 指标见表 2，“—”表示数据集太大，无法聚类。

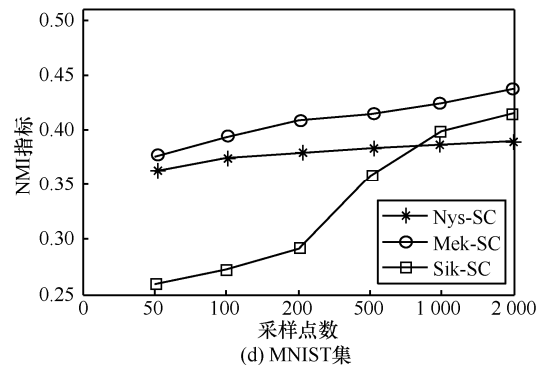
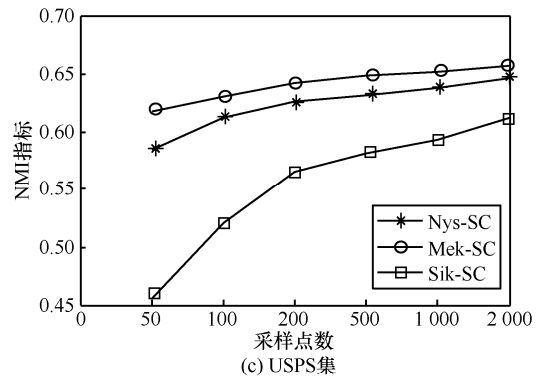
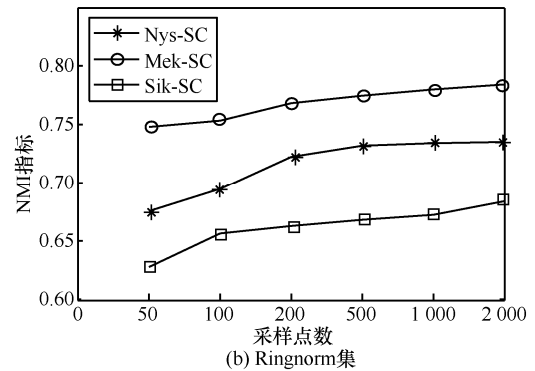
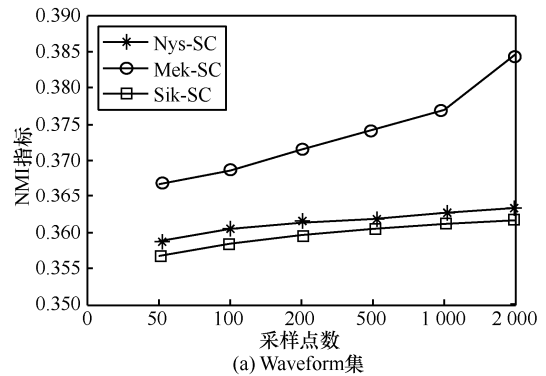


图 1 算法在各数据集上的聚类性能实验结果

表 2 结果显示，Clas-SC 与 Swk-SC 两种算法在 Waveform 集、Ringnorm 集和 USPS 集 3 个数据规模相对较小的数据集取得了较好的聚类结果，



表2 Clas-SC 与 Swk-SC 实验结果

算法	数据集			
	Waveform	Ringnorm	USPS	MNIST
Clas-SC	0.375 8	0.787 4	0.669 5	-
Swk-SC	0.389 5	0.793 2	0.685 5	-

但应用于 MNIST 大规模数据集时无法聚类, 主要因为两种算法完整的核矩阵占用资源巨大, 而从图 1 实验结果可以看出, 通过数据抽样, Nys-SC、Mek-SC 以及 Sik-SC 算法在 4 个数据集上都取得较好的聚类结果, 说明通过数据抽样降低算法资源占用以解决大数据谱聚类是有效的。将图 1 结果与表 2 实验结果进行对比分析, 随着采样点数的增加, Sik-SC 算法的 NMI 指标逐渐逼近 Swk-SC 算法, 说明其近似核矩阵取得与完整核矩阵相近的聚类结果, 也说明改进方法有效。

5.2 算法运行时间比较实验

如图 2 所示为 3 种算法在 4 个数据集上不同采样点数时的运行时间比较结果, Clas-SC 与 Swk-SC 的运行时间见表 3。

表3 Clas-SC 与 Swk-SC 运行时间实验结果

算法	数据集			
	Waveform	Ringnorm	USPS	MNIST
Clas-SC	247.484 2	707.356 8	1 835.890 9	-
Swk-SC	157.670 1	428.077 7	1 164.494 3	-

从实验结果可以看出, Clas-SC 的运行时间最长, 主要因为其特征矩阵特征分解占用很多资源, Swk-SC 采用加权核 k -means 替代矩阵特征分解, 聚类效率有了明显的提高, 但其全部核矩阵的使用在数据规模巨大时仍受限, 而仅使用部分核矩阵的 Nys-SC、Mek-SC 和 Sik-SC 算法的运行时间大幅缩减, 且在 MNIST 集上仍保持较快的聚类运行时间。从图 2 可以看出, 随着采样点数的增加, 3 种算法的聚类时间也逐渐增加, 但 Sik-SC 算法的变化趋势相对最为缓慢, 且大多数情况下聚类时间较短, 这主要因为, Nys-SC 算法的近似特征向量的计算时间随采样数大幅增加, 而 Mek-SC 算法的最佳 k 秩近似核矩阵的求解效率也会大幅降低, Sik-SC 算法仅与采样点数及迭代次数相关,

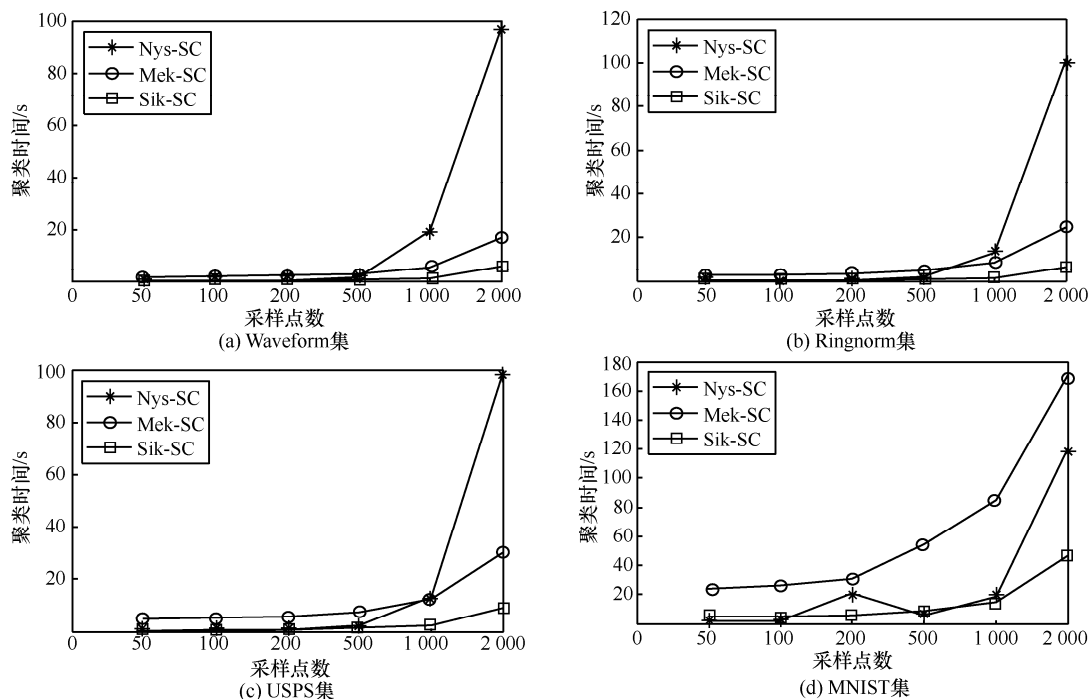


图2 算法在各数据集上的运行时间比较实验结果

聚类效率更高。

综合实验结果, Sik-SC 算法在保持与经典算法聚类性能相似的情况下, 极大地提高了运算效率, 更适合大规模数据挖掘工作。

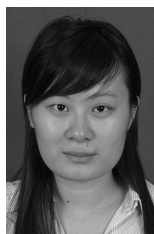
6 结束语

在分析经典谱聚类算法的 Normalized Cut 目标函数与加权 k -means 函数等价的基础上, 设计了一种抽样子空间约束的改进大规模数据谱聚类算法, 算法通过迭代优化避免矩阵特征分解资源占用, 通过将聚类中心约束到抽样子空间中, 避免全部核矩阵的使用, 从而降低时间空间复杂度。对比实验结果验证了算法的有效性。但是改进算法还需进一步考虑如何合理设置采样数以提高算法的自适应性能。

参考文献:

- [1] 成宝芝, 赵春晖, 张丽丽, 等. 联合空间预处理与谱聚类的协同稀疏高光谱异常检测[J]. 光学学报, 2017, 37(4): 296-306.
CHENG B Z, ZHAO C H, ZHANG L L, et al. Cooperative sparse hyperspectral anomaly detection based on joint spatial preprocessing and spectral clustering[J]. Acta Optica Sinica, 2017, 37(4): 296-306.
- [2] 林顺富, 田二伟, 符杨, 等. 基于信息熵分段聚合近似和谱聚类的负荷分类方法[J]. 中国电机工程学报, 2017, 37(8): 2242-2252.
LIN S F, TIAN E W, FU Y, et al. Load classification method based on information entropy segmentation aggregation approximation and spectral clustering[J]. Proceedings of the CSEE, 2017, 37(8): 2242-2252.
- [3] 刘春, 邹海峰, 向勇. 大数据环境下电信数据服务能力开放研究[J]. 电信科学, 2014, 30(3): 156-161.
LIU C, ZOU H F, XIANG Y. Research on telecom data service open ability under the environment of big data[J]. Telecommunications Science, 2014, 30(3): 156-161.
- [4] YAN D H, HUANG L, JORDAN M I. Fast approximate spectral clustering[C]//The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 1, 2009, Paris, France. New York: ACM Press, 2009: 907-916.
- [5] 钱鹏江, 王士同, 邓赵红, 等. 基于最小包含球的大数据集快速谱聚类算法[J]. 电子学报, 2010, 38(9): 2035-2041.
QIAN P J, WANG S T, DENG Z H, et al. Fast spectral clustering for large data sets using minimal enclosing ball[J]. Acta Electronica Sinica, 2010, 38(9): 2035-2041.
- [6] 杨艺, 马儒宁. 基于核心点的大数据谱聚类算法[J]. 中国科学技术大学学报, 2016, 46(9): 757-763.
YANG Y, MA R N. Big data spectrum clustering algorithm based on core points[J]. Journal of University of Science and Technology of China, 2016, 46(9): 757-763.
- [7] KUMAR S, MOHRI M, TALWALKAR A. Sampling methods for the Nyström method[J]. Journal of Machine Learning Research, 2012, 13(1): 981-1006.
- [8] SONG Y Q, CHEN W Y, BAI H J, et al. Parallel spectral clustering[C]//European Conference on Machine Learning and Knowledge Discovery in Databases, September 15-19, 2008, Antwerp, Belgium. New York: ACM Press, 2008: 374-389.
- [9] SONG Y Q, CHEN W Y, BAI H J, et al. Parallel spectral clustering in distributed systems[J]. IEEE Transactions on Pattern Analysis & Mach, 2011, 33(3): 568-586.
- [10] DHILLON I S, GUAN Y, KULIS B. Weighted graph cuts without eigenvectors: a multilevel approach[J]. IEEE Transactions on Pattern Analysis & Mach, 2007, 29(11): 1944-1957.
- [11] HE L, ZHANG H. Kernel K-means sampling for Nyström approximation[J]. IEEE Trans Image Process, 2018, 27(5): 2108-2120.
- [12] 朱光辉, 黄圣彬, 袁春风, 等. SCoS: 基于 Spark 的并行谱聚类算法设计与实现[J]. 计算机学报, 2018, 41(4): 868-884.
ZHU G H, HUANG S B, YUAN C F, et al. SCoS: design and implementation of parallel spectrum clustering algorithm based on Spark[J]. Chinese Journal of Computers, 2018, 41(4): 868-884.
- [13] BÜHLER T, HEIN M. Spectral clustering based on the graph p-Laplacian[C]//The 26th Annual International Conference on Machine Learning, June 14-18, 2009, Montreal, Quebec, Canada. New York: ACM Press, 2009: 81-88.
- [14] 夏景明, 唐玲玲, 谈玲, 等. 基于 K-means 和 MTL-SVM 算法的生理参数监测系统[J]. 电信科学, 2017, 33(10): 43-49.
XIA J, TANG L, TAN L, et al. Biometric monitoring system based on K-means & MTL-SVM algorithm[J]. Telecommunications Science, 2017, 33(10): 43-49.
- [15] CAI D, HE X F, HAN J W, et al. Graph regularized nonnegative matrix factorization for data representation[J]. IEEE Trans Pattern Anal Mach Intell, 2011, 33(8): 1548-1560.
- [16] HAVENS T C, BEZDEK J C, LECKIE C, et al. Fuzzy c-means algorithms for very large data[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(6): 1130-1146.

作者简介



聂茹 (1982-), 女, 华南理工大学广州学院电子信息工程学院讲师、教学研究室副主任, 主要研究方向为智能信息处理、智能控制、电力电子与自动控制等。