软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2037−2049 [doi: 10.13328/j.cnki.jos.004643]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于自适应 Nyström 采样的大数据谱聚类算法*

丁世飞[1,2], 贾洪杰[1,2], 史忠植[2]

[1](中国矿业大学 计算机科学与技术学院,江苏 徐州 221116)

[2](中国科学院 计算技术研究所 智能信息处理重点实验室,北京 100190)

通讯作者: 丁世飞, E-mail: dingsf@cumt.edu.cn

摘 要: 面对结构复杂的数据集,谱聚类是一种灵活而有效的聚类方法,它基于谱图理论,通过将数据点映射到一个由特征向量构成的低维空间,优化数据的结构,得到令人满意的聚类结果.但在谱聚类的过程中,特征分解的计算复杂度通常为 $O(n^3)$,限制了谱聚类算法在大数据中的应用.Nyström 扩展方法利用数据集中的部分抽样点,进行近似计算,逼近真实的特征空间,可以有效降低计算复杂度,为大数据谱聚类算法提供了新思路.抽样策略的选择对 Nyström 扩展技术至关重要,设计了一种自适应的 Nyström 采样方法,每个数据点的抽样概率都会在一次采样完成后及时更新,而且从理论上证明了抽样误差会随着采样次数的增加呈指数下降.基于自适应的 Nyström 采样方法,提出一种适用于大数据的谱聚类算法,并对该算法的可行性和有效性进行了实验验证.

关键词: 大数据;谱聚类;特征分解;Nyström 扩展;自适应采样

中图法分类号: TP181

中文引用格式: 丁世飞,贾洪杰,史忠植.基于自适应 Nyström 采样的大数据谱聚类算法.软件学报,2014,25(9):2037−2049. http://www.jos.org.cn/1000-9825/4643.htm

英文引用格式: Ding SF, Jia HJ, Shi ZZ. Spectral clustering algorithm based on adaptive Nyström sampling for big data analysis. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2037−2049 (in Chinese). http://www.jos.org.cn/1000-9825/4643.htm

Spectral Clustering Algorithm Based on Adaptive Nyström Sampling for Big Data Analysis

DING Shi-Fei[1,2], JIA Hong-Jie[1,2], SHI Zhong-Zhi[2]

[1](School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

[2](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

Corresponding author: DING Shi-Fei, E-mail: dingsf@cumt.edu.cn

Abstract: Spectral clustering is a flexible and effective clustering method for complex structure data sets. It is based on spectral graph theory and can produce satisfactory clustering results by mapping the data points into a low-dimensional space constituted by eigenvectors so that the data structure is optimized. But in the process of spectral clustering, the computational complexity of eigen-decomposition is usually $O(n^3)$, which limits the application of spectral clustering algorithm in big data problems. Nyström extension method uses partial points sampled from the data set and approximate calculation to simulate the real eigenspace. In this way, the computational complexity can be effectively reduced, which provides a new idea for big data spectral clustering algorithm. The selection of sampling strategy is essential for Nyström extension technology. In this paper, the design of an adaptive Nyström sampling method is presented. The sampling probability of every data point will be updated after each sampling pass, and a proof is given that the sampling error will decrease exponentially with the increase of sample times. Based on the adaptive Nyström sampling method, a spectral clustering algorithm for big data analysis is presented, and its feasibility and effectiveness is verified by experiments.

Key words: big data; spectral clustering; eigen-decomposition; Nyström extension; adaptive sampling

* 基金项目: 国家重点基础研究发展计划(973)(2013CB329502); 国家自然科学基金(61379101)
收稿时间: 2014-04-07; 定稿时间: 2014-05-14

聚类学习是一种重要的数据分析技术.为了从纷繁复杂的数据中发现有用的信息,可以先对数据进行聚类,根据数据对象的相关特征,将相似的对象归到同一类里,而差别较大的对象划分到不同类中,找到数据之间的内在联系,为决策提供支持[1].谱聚类是聚类分析中十分热门的研究领域,与传统的聚类算法(如 k-means, FCM)相比,其优势在于:谱聚类算法可以很好地处理非凸形结构的数据集,得到比较满意的聚类结果[2].谱聚类的背后有着坚实的理论基础,它用图划分的思想处理数据聚类问题,为了得到最优的子图划分,引入拉普拉斯矩阵并对其特征分解,利用特征向量将原始数据点映射到一个低维的特征空间中,再进行聚类.

但是传统的谱聚类算法只适用于规模较小的数据集,因为在其聚类过程中,存储相似度矩阵需要的空间复杂度是 $O(n^2)$;而对拉普拉斯矩阵特征分解,需要的时间复杂度一般为 $O(n^3)$,这样的复杂度在处理大规模数据时是无法接受的[3].所以,如何降低谱聚类算法的计算复杂度,将其应用于大数据的聚类上,一直是个非常有挑战性的难题.近年来,该问题的研究受到了越来越多的关注.Song 等人[4]设计了一种并行的谱聚类算法,将矩阵稀疏化处理,同时利用机器集群的优势对大规模的数据集进行聚类.Yan 等人[5]从另外的思路,提出了一种快速逼近的谱聚类算法,首先,通过 k-means 或 RP-tree 算法对数据集进行失真最小化局部变换,得到代表点的集合;然后,利用谱聚类算法将这些代表点划分成若干类;最后,根据所有点与代表点的一致性,恢复所有点的类属关系.经典的谱聚类算法一般选择前 k 个特征向量构成低维嵌入空间,但是 Lin 等人[6]发现特征向量具有收敛性,利用幂方法对其多次迭代,得到的第 1 个特征向量具备多路聚类所需要的信息,从而降低了算法复杂度.

当前,在谱聚类研究领域,Nyström 扩展技术是一种有效的用于大规模数据聚类的方法.Nyström 最初的设计是为了求解积分方程[7],其原理是:利用少量的抽样点对连续或离散空间中的卷积算子进行逼近,求解近似的特征向量.Williams 和 Seeger[8]将 Nyström 方法引入到核机器学习中,以加速 Gram 矩阵的特征分解,而准确率并没有明显下降.Fowlkes 等人[9]扩展了 Nyström 方法,使其可以处理规范割(normalized cut)问题,并用改进的谱聚类算法进行图像分割,取得了很好的效果.

一般情况下,Nyström 扩展方法的性能与样本点的选取有很大关系.通常认为,抽样的数目越多,Nyström 逼近的效果越好,得到的计算结果与真实值的差别越小[10].但是大量抽样也有明显的弊端:一方面没有抽样停止的标准,算法无休止地抽样是不可能的;另一方面,随着抽样点的增多,算法的复杂度必然大幅度增加.最初用于谱聚类的 Nyström 扩展技术是通过随机抽样获得样本点,然而随机抽样具有不稳定性,在处理大数据集时,样本点容易集中在某一局部区域,产生不准确的计算结果.因此,采用何种策略来抽样,使样本点更好地反映原数据集的分布情况,从而降低逼近误差,是 Nyström 方法的一个十分关键的问题.Zhang 等人[11,12]详细分析了 Nyström 扩展的低秩逼近误差,指出量化误差的目标函数可由 k-means 算法求解,然后将 k-means 初步聚类的中心点作为抽样点,用于 Nyström 扩展方法.Wang 等人[13]提出一种基于最远最近策略的抽样方法,首先根据每个点被抽样的概率,形成 h 组数据.再从这些组中采样,获得最终的样本点集合.本文设计了一种自适应的 Nyström 采样方法,通过多次遍历并更新抽样概率,选取更有意义的样本点,而且从理论上证明了抽样误差会随着遍历次数的增加呈指数下降.基于此,提出一种适用于大数据的基于自适应 Nyström 采样的谱聚类算法,该算法可以用较小的抽样集获得令人满意的聚类效果.

本文第 1 节从图论的角度介绍谱聚类算法,重点分析规范割的基本原理.第 2 节阐述 Nyström 方法的相关理论,以及如何将其扩展并应用于谱聚类算法.第 3 节给出自适应的 Nyström 采样方法,并且从理论层面证明该方法的有效性.第 4 节提出基于自适应 Nyström 采样的谱聚类算法.第 5 节设计实验,将所提出的算法与其他算法进行对比.最后总结本文所做的工作,并给出下一步的研究方向.

1 谱聚类算法原理

给定一个包含 N 个点的数据集,根据数据点的成对相似性可以构造一个 $N\times N$ 的相似性矩阵,谱方法就是基于相似矩阵的特征向量和特征值来聚类的.利用这些特征向量,可以构造数据点的一个低维嵌入子空间,在这个嵌入空间中,可以使用简单的中心聚类方法(例如 k-means)对数据点进行聚类.规范割(normalized cut)[14]是一个典型的谱聚类算法,下面简要介绍该方法的基本原理.

谱方法是基于图论的,首先将数据点想象成无向加权图 $G(V,E)$,节点 V 代表数据点;边 E 的权重表示数据点的成对相似性,这些相似性的值就形成了对称矩阵 $W\in\mathbb{R}^{N\times N}$.设 A 和 B 分别表示 V 的二部划分,即:$A\cup B=V, A\cap B=\varnothing$.令 $cut(A,B)$表示 A 和 B 之间的权重之和,即:$cut(A,B)=\sum_{i\in A,j\in B}W_{ij}$.第 i 个节点的度定义为 $d_i=\sum_j W_{ij}$,集合的容量(volume)为该集合内所有节点度的总和:$vol(A)=\sum_{i\in A}d_i, vol(B)=\sum_{i\in B}d_i$.集合 A 和 B 之间的规范割由下式给出:

$$NCut(A,B)=\frac{cut(A,B)}{vol(A)}+\frac{cut(A,B)}{vol(B)}.$$

我们希望找到 A 和 B,使 $NCut(A,B)$最小化.借助谱图理论,Shi 和 Malik[14]指出:通过求解归一化拉普拉斯矩阵 L 的第二小特征值λ_2,并对相应的特征向量阈值处理,可以获得该问题的一个近似解.归一化的拉普拉斯矩阵定义为

$$L=D^{-1/2}(D-W)D^{-1/2}=I-D^{-1/2}WD^{-1/2},$$

其中,D 是对角矩阵,其元素 $D_{ii}=d_i$.不管 W 如何,矩阵 L 是半正定的.它的特征值在区间[0,2]内,所以 $D^{-1/2}WD^{-1/2}$ 的特征值被限制在[−1,1]中.

扩展到多分类问题,可以通过递归二划分或使用多个特征向量来处理[15].本文采用多个特征向量把每个元素嵌入到一个 N_E-维欧氏空间($N_E<<N$),以便保留归一化相似性中的显著差异,抑制其中的噪音.然后,使用 k-means 算法对嵌入空间中的点进行划分.

要找到这样的嵌入空间,需要对 $D^{-1/2}WD^{-1/2}$ 特征分解,计算其特征向量和特征值:

$$(D^{-1/2}WD^{-1/2})V=V\Lambda,$$

其中,V 由特征向量组成,是一个 $N\times N_E$ 的矩阵;Λ由特征值组成,是 $N_E\times N_E$ 的对角矩阵.第 j 个点的第 i 个嵌入坐标为

$$E_{ij}=\frac{V_{i+1,j}}{\sqrt{D_{jj}}}, i=1,...,N_E, j=1,...,N,$$

其中,特征向量已经按特征值的升序排列.

但是,解决这个问题所需的计算量是非常大的.在聚类的过程中,W 以元素数二次幂的规模增长,很快就会占满内存空间,更不用说计算其特征向量了.一种解决方法是使用稀疏、近似的 W,其中每个元素仅与其附近的少数邻居相连,而其他连接都假定为 0[16].这样就可以采用高效的、用于稀疏矩阵的特征值求解方法(例如 Lanczos 法),不过该方法的有效性还有待进一步研究.Fowlkes 等人[9]提出的基于 Nyström 矩阵低秩逼近的解决方案,允许保留所有的相似性值,虽然也会损失一部分精度,但是极大地降低了计算复杂度,在实践中取得了较好效果.

2 Nyström 扩展技术

Nyström 方法是寻找数值近似的技术,适用于如下特征函数问题:

$$\int_a^b W(x,y)\phi(y)\mathrm{d}y=\lambda\phi(x).$$

我们可以在区间[a,b]上选取一组均匀分布的点$\xi_1,\xi_2,\ldots,\xi_n$,然后用简单的求积公式来逼近该积分方程:

$$\frac{(b-a)}{n}\sum_{j=1}^{n}W(x,\xi_j)\hat{\phi}(\xi_j)=\lambda\hat{\phi}(x) \tag{1}$$

其中,$\hat{\phi}(x)$ 是真实$\phi(x)$的一个近似.

由于公式(1)对任意 x 都成立,为了求解,可以用抽样点代替 x,即:令 $x=\xi_i$,得到下式:

$$\frac{(b-a)}{n}\sum_{j=1}^{n}W(\xi_i,\xi_j)\hat{\phi}(\xi_j)=\lambda\hat{\phi}(\xi_i),\forall i\in\{1,...,n\}.$$

不失一般性,将[a,b]替换成[0,1],上述问题等价于矩阵特征值问题:

$$A\hat{\Phi} = n\hat{\Phi}\Lambda,$$

其中,$A_{ij}=W(\xi_i,\xi_j)$;$\Phi=[\phi_1,\phi_2,\ldots,\phi_n]$是 A 的 n 个特征向量,相应的特征值为$\lambda_1,\lambda_2,\ldots,\lambda_n$.代入到公式(1)中,可以得到每个 $\hat{\phi}_i$ 的 Nyström 扩展:

$$\hat{\phi}_i(x) = \frac{1}{n\lambda_i}\sum_{j=1}^{n} W(x,\xi_j)\hat{\phi}_i(\xi_j) \tag{2}$$

公式(2)可以将一组样本点的特征向量扩展到任意使用 $W(\cdot,\xi_j)$作为插值权重的点 x,尽管 x 可以是任意真值,我们所关心的是如何处理那些没有被抽到的点.为了便于理解,下面从矩阵补全的角度来分析 Nyström 扩展的性质.

将所有 N 个数据点分成两部分,一部分为随机抽样得到的 n 个样本点,另一部分为剩余的 N-n 个数据点,则相似矩阵 W 可以写成:

$$W = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{3}$$

其中,$A\in\mathbb{R}^{n\times n}$ 为抽样点间的相似度矩阵,且 $A=U\Lambda U^T$;$B\in\mathbb{R}^{(N-n)\times n}$ 为抽样点与剩余点的相似度矩阵;$C\in\mathbb{R}^{(N-n)\times(N-n)}$ 为剩余点间的相似度矩阵.令 $\bar{U}$ 表示 W 的近似特征向量,由 Nyström 扩展可以得到:

$$\bar{U} = \begin{bmatrix} U \\ B^T U\Lambda^{-1} \end{bmatrix}.$$

相应地,令 $\hat{W}$ 表示近似的 W,则有:

$$\hat{W} = \bar{U}\Lambda\bar{U}^T = \begin{bmatrix} U \\ B^T U\Lambda^{-1} \end{bmatrix}\Lambda[U^T \quad \Lambda^{-1}U^T B] = \begin{bmatrix} U\Lambda U^T & B \\ B^T & B^T A^{-1}B \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & B^T A^{-1}B \end{bmatrix}.$$

可见,Nyström 扩展技术使用 $B^TA^{-1}B$ 来逼近 C.由于 $n<<N$,所以抽样后剩余点的数目通常是很大的,经过 Nyström 逼近,避免了使用剩余点的相似度,从而极大地降低了问题求解的空间和时间复杂度.但近似特征向量 $\bar{U}$ 并不能直接使用,因为不一定满足特征向量正交的性质.定理 1 给出了 $\hat{W}$ 正交特征向量的表达式.

定理 1. 若 A 是正定的,令 $A^{1/2}$ 表示 A 的对称正定平方根,定义 $Q=A+A^{-1/2}BB^TA^{-1/2}$,将其对角化 $Q=U_Q\Lambda_Q U_Q^T$,则 $\hat{W}$ 的正交特征向量为

$$V = \begin{bmatrix} A \\ B^T \end{bmatrix}A^{-1/2}U_Q\Lambda_Q^{-1/2} \tag{4}$$

证明:

(1) 首先证明 V 是 $\hat{W}$ 的特征向量:

$$\hat{W} = \begin{bmatrix} A & B \\ B^T & B^TA^{-1}B \end{bmatrix} = \begin{bmatrix} A \\ B^T \end{bmatrix}A^{-1}[A \quad B] = \left\{\begin{bmatrix} A \\ B^T \end{bmatrix}A^{-1/2}U_Q\Lambda_Q^{-1/2}\right\}\Lambda_Q\{\Lambda_Q^{-1/2}U_Q^TA^{-1/2}[A \quad B]\} = V\Lambda_QV^T.$$

(2) 然后证明 V^T 和 V 是正交的:

$$I = V^TV = \{\Lambda_Q^{-1/2}U_Q^TA^{-1/2}[A \quad B]\}\left\{\begin{bmatrix} A \\ B^T \end{bmatrix}A^{-1/2}U_Q\Lambda_Q^{-1/2}\right\}.$$

上式右边的两部分分别乘以 $U_Q\Lambda_Q^{1/2}$ 和 $\Lambda_Q^{1/2}U_Q^T$,即可得到 Q:

$$Q = U_Q\Lambda_QU_Q^T = A^{-1/2}[A \quad B]\begin{bmatrix} A \\ B^T \end{bmatrix}A^{-1/2} = A + A^{-1/2}BB^TA^{-1/2}.$$ □

要将 Nyström 逼近用于谱聚类,还必须对相似矩阵(拉普拉斯矩阵)归一化处理,即 $\hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2}$, 其中, $\hat{D}$ 是对角矩阵,其对角线元素 $\hat{D}_{ii}$ 等于 $\hat{W}$ 第 i 行元素之和.Fowlkes 等人[9]给出了节点度的一种简便的计算方法:

$$\hat{\boldsymbol{d}} = \hat{W}\mathbf{1} = \begin{bmatrix} A\mathbf{1}_n + B\mathbf{1}_m \\ B^T\mathbf{1}_n + B^TA^{-1}B\mathbf{1}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_r + \boldsymbol{b}_r \\ \boldsymbol{b}_c + B^TA^{-1}\boldsymbol{b}_r \end{bmatrix} \tag{5}$$

其中,$m=N-n$,$\boldsymbol{a}_r,\boldsymbol{b}_r\in\mathbb{R}^m$ 分别表示 A 和 B 的行和,$\boldsymbol{b}_c\in\mathbb{R}^n$ 表示 B 的列和,$\mathbf{1}$ 表示元素均为 1 的列向量.利用 $\hat{\boldsymbol{d}}$ 就可以

将矩阵 A 和 B 归一化:

$$A_{ij} \leftarrow \frac{A_{ij}}{\sqrt{\hat{\boldsymbol{d}}_i \hat{\boldsymbol{d}}_j}}, i,j=1,...,n \tag{6}$$

$$B_{ij} \leftarrow \frac{B_{ij}}{\sqrt{\hat{\boldsymbol{d}}_i \hat{\boldsymbol{d}}_{j+n}}}, i=1,...,n, j=1,...,m \tag{7}$$

3 自适应的 Nyström 采样方法

Nyström 扩展技术不但易于理解和实现,而且已经应用到很多有挑战性的场景中,取得了令人满意的效果.Nyström 方法的关键步骤是,从原数据集中选择一定数量的样本点.最常想到的方式是采用随机抽样,但是随机抽样具有不稳定性,而且对于大规模的数据集,很难做到均匀采样,所抽的数据点可能集中在某一局部区域,从而导致较差的聚类结果.

谱聚类的基本思想是:求解相似矩阵的前 k 个特征向量,将原始数据点映射到由特征向量构成的一个低维子空间中;在这样的近似空间中,数据的结构得到了优化,同一类中的点会更加紧凑,而不同类中的点会更加分离,聚类的结果也会更好.Frieze 等人[17]指出:任何矩阵 A 都有 k/ε行,其生成空间可以形成 A 的一个秩-k 近似,附加的误差在 $\varepsilon \| A \|_F^2$ 范围内.而所选取行的子集可以作为独立的样本,从 A 各行范数决定的分布中获得.

定理 2. 给定一个 $m\times n$ 的矩阵 A,从中选择 s 行构造矩阵 S,每次采样都独立地遵循下面的分布,第 i 行被选取的概率为

$$P_i = \frac{\| A^{(i)} \|^2}{\| A \|_F^2},$$

其中,$A^{(i)}$表示 A 的第 i 行.如果 $s \geqslant k/\varepsilon$,矩阵 $\tilde{A}_k$ (秩最大为 k)的行就包含在 $span(S)$中,且满足:

$$\mathbb{E}_S(\| A-\tilde{A}_k \|_F^2) \leqslant \| A-A_k \|_F^2 + \varepsilon \| A \|_F^2 .$$

定理 2 可以转化为高效的采样算法[18],该算法首先遍历一次 A 得到样本的概率分布,然后根据每一行的概率进行采样和近似计算,其复杂度是 $O(\min\{m,n\}k^2/\varepsilon^4)$.由此考虑,逼近误差是否可以通过多次遍历数据而显著减少?举例来说,假设数据中除了一个点之外,其余的点都沿着$\mathbb{R}^n$ 的一个 1 维子空间,最好的秩-2 子空间应该是零误差的.然而,一轮采样很有可能遗漏远离线的这一点.所以想到用两轮采样的方法:在第 1 轮,首先从平方范的分布中得到一组样本;然后,重新计算剩余点的抽样概率,每个点被选取的概率正比于其与已有样本生成空间的平方距离,根据概率的大小选择另一组样本,这个过程称为自适应采样.如果孤立点错过了第 1 次采样,它在第 2 次采样中被选择的概率就会很高.现在全部样本的生成空间就能形成一个良好的秩-2 近似.可以证明:当进行自适应地采样时,附加误差随着遍历次数的增加呈指数下降.因此,多次遍历数据对低秩近似问题是非常有益的.自适应采样的数学描述由定理 3 给出.

定理 3. 令 $S=S_1\cup\ldots\cup S_t$ 是对 $m\times n$ 矩阵 A 的行的随机抽样,其中对于 j=1,…,t,每个集合 S_j 都是 A 的 s 行的一个采样,从以下分布中独立地选择:行 i 被选中的概率为

$$P_i^{(j)} = \frac{\| E_j^{(i)} \|^2}{\| E_j \|_F^2},$$

其中, $E_1 = A, E_j = A - \pi_{S_1\cup\ldots\cup S_{j-1}}(A)$,$\pi_S(A)$ 表示 A 在 S 的生成空间上的投影.对于 $s \geqslant k/\varepsilon$,矩阵 $\tilde{A}_k$ (秩最大为 k)的行就包含在 $span(S)$中,且满足:

$$\mathbb{E}_S(\| A-\tilde{A}_k \|_F^2) \leqslant \frac{1}{1-\varepsilon}\| A-A_k \|_F^2 + \varepsilon^t \| A \|_F^2 .$$

从定理 3 可知,虽然抽样分布改变了 t 次,矩阵本身没有改变.这样,当 A 为稀疏矩阵时就显得特别重要.由于保持了矩阵的稀疏性,可以充分利用稀疏性减少计算量,提高程序的运行效率.定理 3 的证明将在第 3.2 节给出,在证明前,先回顾矩阵的相关知识.

3.1 预备知识

任意 $m\times n$ 的实矩阵 A 都可以进行奇异值分解,即矩阵 A 可以写成下面的形式:

$$A=\sum_{i=1}^{r}\sigma_i u^{(i)}v^{(i)^T},$$

其中,r 是 A 的秩,$\sigma_1\geqslant\sigma_2\geqslant\ldots\geqslant\sigma_r\geqslant 0$ 称为奇异值;$\{u^{(1)},\ldots,u^{(r)}\}\in\mathbb{R}^m$,$\{v^{(1)},\ldots,v^{(r)}\}\in\mathbb{R}^n$ 是正交向量的集合,它们分别称为左奇异向量和右奇异向量,遵循 $A^Tu^{(i)}=\sigma_i v^{(i)}$ 和 $A^Tv^{(i)}=\sigma_i u^{(i)}$,其中,$1\leqslant i\leqslant r$.

矩阵 $A\in\mathbb{R}^{m\times n}$ 的元素为 a_{ij},A 的 Frobenius 范数记作 $\|A\|_F$,由下式给出:

$$\|A\|_F^2=\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2=\sum_{i=1}^{r}\sigma_i^2.$$

对于子空间 $V\subseteq\mathbb{R}^n$,令 $\pi_{V,k}(A)$ 表示 A 的最佳的秩-k 近似(在 Frobenius 范数下),V 中元素为 A 的行索引.令:

$$\pi_k(A)=\pi_{\mathbb{R}^n,k}(A)=\sum_{i=1}^{k}\sigma_i u^{(i)}v^{(i)^T}$$

为 A 的最佳的秩-k 近似.并且 $\pi_V(A)=\pi_{V,n}(A)$ 是 A 在 V 上的正交投影.我们说"A 的一组行的集合(或样本)",意思是一组行的索引,而不是实际的行.对于 A 的一组行的集合 S,令 $span(S)\subseteq\mathbb{R}^n$ 表示这些行产生的子空间;为描述方便,$\pi_{span(S)}(A)$ 简记为 $\pi_S(A)$,$\pi_{span(S),k}(A)$ 简记为 $\pi_{S,k}(A)$.

对于子空间 $V,W\subseteq\mathbb{R}^n$,它们的和记作 $V+W$,由下式给出:

$$V+W=\{x+y\in\mathbb{R}^n:x\in V,y\in W\}.$$

同时,还会用到操作符 π_V 的以下基本性质:

(1) π_V 是线性的,即 $\pi_V(\lambda A+B)=\lambda\pi_V(A)+\pi_V(B)$,对任意 $\lambda\in\mathbb{R}$,矩阵 $A,B\in\mathbb{R}^{m\times n}$ 成立;

(2) 如果 $V,W\in\mathbb{R}^n$ 是正交线性子空间,则 $\pi_{V+W}(A)=\pi_V(A)+\pi_W(A)$,对任意矩阵 $A\in\mathbb{R}^{m\times n}$ 成立.

对于一个随机向量 v,其期望记作 $E(v)$.$E(v)$ 也是一个向量,其中的每个元素都是 v 中元素的预期值.

3.2 自适应采样的证明

为了便于证明,定义一个中间引理如下.

引理 1. 令 $A\in\mathbb{R}^{m\times n}$,$V\subseteq\mathbb{R}^n$ 是一个向量子空间.令 $E=A-\pi_V(A)$,S 为包含 s 行的随机抽样,每一行都从 A 中选取,选取的概率符合分布 D:

$$P_i=\frac{\|E^{(i)}\|^2}{\|E\|_F^2}\tag{8}$$

然后,对于任何非负整数 k,都有:

$$\mathbb{E}_S(\|A-\pi_{V+span(S),k}(A)\|_F^2)\leqslant\|A-\pi_k(A)\|_F^2+\frac{k}{s}\|E\|_F^2.$$

证明:定义向量 $w^{(1)},\ldots,w^{(k)}\in V+span(S)$,令 $W=span\{w^{(1)},\ldots,w^{(k)}\}$,则 W 可以认为是 $span\{v^{(1)},\ldots,v^{(k)}\}$ 的一个很好的近似,即:

$$\mathbb{E}_S(\|A-\pi_W(A)\|_F^2)\leqslant\|A-\pi_k(A)\|_F^2+\frac{k}{s}\|E\|_F^2\tag{9}$$

其中,$\pi_k(A)=\pi_{span\{v^{(1)},\ldots,v^{(k)}\}}(A)$,$span\{v^{(1)},\ldots,v^{(k)}\}$ 表示投影的最佳子空间.证明了公式(9)就证明了引理 1,因为 $W\subseteq V+span(S)$.

为此,定义 $X_l^{(j)}$ 为一个随机变量,其概率为 P_i,使得对于 $i=1,\ldots,m$ 和 $l=1,\ldots,s$,有:

$$X_l^{(j)}=\frac{u_i^{(j)}}{P_i}E^{(i)}=\frac{u_i^{(j)}}{P_i}(A^{(i)}-\pi_V(A^{(i)})).$$

注意到 $X_l^{(j)}$ 是 A 中某一行的线性函数,该行从分布 D 中采样得到.令 $X^{(j)}=\frac{1}{s}\sum_{l=1}^{s}X_l^{(j)}$,则

$$\mathbb{E}_S(X^{(j)})=E^Tu^{(j)}.$$

对于 $1\leqslant j\leqslant k$,定义:

$$w^{(j)})=\pi_W(A)^Tu^{(j)}+X^{(j)} \tag{10}$$

然后可以得到$\mathbb{E}_S(w^{(j)})=\sigma_jv^{(j)}$.下一步要寻找约束 $w^{(j)}$的二阶中心矩,即$\mathbb{E}_S(\|w^{(j)}-\sigma_jv^{(j)}\|^2)$.因为

$$w^{(j)}-\sigma_jv^{(j)}=X^{(j)}-E^Tu^{(j)},$$

所以,

$$\begin{aligned}\mathbb{E}_s(\|w^{(j)}-\sigma_jv^{(j)}\|^2)&=\mathbb{E}_S(\|X^{(j)}-E^Tu^{(j)}\|^2)\\&=\mathbb{E}_S(\|X^{(j)}\|^2)-2\mathbb{E}_S(X^{(j)})\cdot E^Tu^{(j)}+\|E^Tu^{(j)}\|^2\\&=\mathbb{E}_S(\|X^{(j)}\|^2)-\|E^Tu^{(j)}\|^2\end{aligned} \tag{11}$$

单独分析公式(11)的第 1 项:

$$\begin{aligned}\mathbb{E}_S(\|X^{(j)}\|^2)&=\mathbb{E}_S\left(\left\|\frac{1}{s}\sum_{l=1}^{s}X_l^{(j)}\right\|^2\right)\\&=\frac{1}{s^2}\sum_{l=1}^{s}\mathbb{E}_S(\|X_l^{(j)}\|^2)+\frac{2}{s^2}\sum_{1\leqslant l_1\leqslant l_2\leqslant s}\mathbb{E}_S(X_{l_1}^{(j)}\cdot X_{l_2}^{(j)})\\&=\frac{1}{s^2}\sum_{l=1}^{s}\mathbb{E}_S(\|X_l^{(j)}\|^2)+\frac{s-1}{s}\|E^Tu^{(j)}\|^2\end{aligned} \tag{12}$$

在公式(12)中, $X_{l_1}^{(j)}$和 $X_{l_2}^{(j)}$是相互独立的.由公式(11)和公式(12)可以得到:

$$\mathbb{E}_S(\|w^{(j)}-\sigma_jv^{(j)}\|^2)=\frac{1}{s^2}\sum_{l=1}^{s}\mathbb{E}_s(\|X_l^{(j)}\|^2)-\frac{1}{s}\|E^Tu^{(j)}\|^2 \tag{13}$$

由 P_i的定义可知:

$$\mathbb{E}_S(\|X_l^{(j)}\|^2)=\sum_{i=1}^{m}P_i\frac{\|u_i^{(j)}E^{(i)}\|}{P_i^2}\leqslant\|E\|_F^2 \tag{14}$$

利用公式(13)和公式(14),就获得了 $w^{(j)}$的二阶中心矩上的一个约束:

$$\mathbb{E}_S(\|w^{(j)}-\sigma_jv^{(j)}\|^2)\leqslant\frac{1}{s}\|E\|_F^2 \tag{15}$$

有了这个约束,即可完成证明.令 $y^{(j)}=w^{(j)}/\sigma_j,1\leqslant j\leqslant k$,定义矩阵 $F=A\sum_{i=1}^{k}v^{(i)}y^{(i)^T}$, 可以用 F 来约束误差 $\|A-\pi_W(A)\|_F^2$.F 的行空间包含在 $W=span\{w^{(1)},\ldots,w^{(k)}\}$中,所以$\|A-\pi_W(A)\|_F^2\leqslant\|A-F\|_F^2$.

通过沿左奇异向量 $u^{(1)},\ldots,u^{(r)}$分解 $A-F$,可以使用不等式(15)来约束$\|A-F\|_F^2$,从而证明引理 1:

$$\begin{aligned}\mathbb{E}_S(\|A-\pi_W(A)\|_F^2)&\leqslant\mathbb{E}_S(\|A-F\|_F^2)\\&=\sum_{i=1}^{r}\mathbb{E}_S(\|(A-F)^Tu^{(i)}\|_F^2)\\&=\sum_{i=1}^{k}\mathbb{E}_S(\|\sigma_iv^{(i)}-w^{(i)}\|^2)+\sum_{i=k+1}^{r}\sigma_i^2\\&\leqslant\frac{k}{s}\|E\|_F^2+\|A-\pi_k(A)\|_F^2\end{aligned} \tag{16}$$

□

利用引理 1,下面采用归纳法证明定理 3.

定理 3 的证明:假设采样的总次数为 t.定理 2 给出了 $t=1$ 的基本情况.

令 $E=A-\pi_{S_1\cup\ldots\cup S_{t-1}}(A)$.借助引理 1,对于 $s\geqslant k/\varepsilon$,可以得到:

$$\mathbb{E}_{S_t}(\| A-\pi_{S_1\cup\ldots\cup S_t,k}(A)\|_F^2)\leqslant\| A-\pi_k(A)\|_F^2+\varepsilon\| E\|_F^2.$$

又因为$\| E\|_F^2\leqslant\| A-\pi_{S_1\cup\cdots\cup S_{t-1},k}(A)\|_F^2$,所以,

$$\mathbb{E}_{S_t}(\| A-\pi_{S_1\cup\ldots\cup S_t,k}(A)\|_F^2)\leqslant\| A-\pi_k(A)\|_F^2+\varepsilon\| A-\pi_{S_1\cup\cdots\cup S_{t-1},k}(A)\|_F^2 \tag{17}$$

利用在$S_1,\ldots,S_{t-1}$上的期望以及t–1次采样的归纳假设,就可以得到定理3的结果:

$$\begin{aligned}\mathbb{E}_S(\| A-\pi_{S_1\cup\ldots\cup S_t,k}(A)\|_F^2)&\leqslant\| A-\pi_k(A)\|_F^2+\varepsilon\mathbb{E}_{S_1,\ldots,S_{t-1}}(\| A-\pi_{S_1\cup\ldots\cup S_{t-1},k}(A)\|_F^2)\\&\leqslant\| A-\pi_k(A)\|_F^2+\varepsilon\left(\frac{1}{1-\varepsilon}\| A-\pi_k(A)\|_F^2+\varepsilon^{t-1}\| A\|_F^2\right)\\&=\frac{1}{1-\varepsilon}\| A-\pi_k(A)\|_F^2+\varepsilon^t\| A\|_F^2\end{aligned} \tag{18}$$

□

当进行t次采样时,所有样本索引的集合为$S_1\cup\ldots\cup S_t=S$.由不等式(18)可知:A在$span(S)$上的二阶中心矩$\mathbb{E}_S(\| A-\pi_{S,k}(A)\|_F^2)$不超过$\frac{1}{1-\varepsilon}\| A-\pi_k(A)\|_F^2+\varepsilon^t\| A\|_F^2$,其中,$\pi_k(A)$是$A$最佳的秩-$k$近似,$\frac{1}{1-\varepsilon}\| A-\pi_k(A)\|_F^2$是一个定值,$\varepsilon^t\| A\|_F^2$是附加的误差.因为0<$\varepsilon$<1,所以当采样次数$t$增大时,附加误差$\varepsilon^t\| A\|_F^2$会呈指数下降.定理3说明,采用自适应的抽样方法可以有效减少抽样误差.通过多次遍历采样,使得到的样本点具有更强的代表性,这样,利用少数抽样点就能较好地描述数据的真实结构.在Nyström扩展方法中,使用这些高质量的样本点进行近似计算,得到相似矩阵的特征向量更接近真实的值.

4 基于自适应Nyström采样的谱聚类算法

第3节从理论层面分析了基于概率分布的自适应抽样方法.传统的算法大多从一个固定分布里对数据点进行采样,而自适应采样每次选择一组样本后,都会更新所有样本的概率分布,使最终得到的样本可以产生最佳的秩-k子空间($v^{(1)},\ldots,v^{(k)}$的生成空间)的一个近似.利用该自适应抽样方法对相似矩阵W进行采样,从所有行的初始分布开始,选择s<n行来形成子矩阵R';然后,根据先前选择的行更新剩余行被选择的概率,选取概率最大的s个新的行并入R'.重复此过程,直到已经选择了n行为止.该自适应采样方案详见算法1.

算法1. 自适应Nyström采样算法.

输入:数据点的相似度矩阵$W\in\mathbb{R}^{N\times N}$,总的采样行数$n$,每次遍历选取的行数$s$;

输出:W中n个行的索引,即抽得的n个样本.

SAMPLE-ADAPTIVE(W,N,n,s)

Step 1. 初始化抽样点的索引的集合$S=\varnothing$,令迭代次数$t=n/s$.

Step 2. 对于$i\in[1,\ldots,t]$,重复执行以下步骤:

(a) $P_i\leftarrow$*UPDATE-PROBABILITY*(S)

(b) $S_i\leftarrow$根据P_i选取的s个索引组成的集合

(c) $S\leftarrow S\cup S_i$

Step 3. 返回抽样集合S

UPDATE-PROBABILITY(S)

Step 1. 选出与S中的索引对应的W的行,组成集合R'.

Step 2. 计算R'的左奇异向量$U_{R'}$

Step 3. 计算W与它在R'上的正交投影的误差:$E=W-U_{R'}U_{R'}^TW$.

Step 4. 对于$j\in[1,\ldots,n]$,重复执行以下步骤:

(a) 如果$j\in S$,就令$P_j=0$;

(b) 否则,令$P_j=\| E_j\|_2^2$.

Step 5. 更新所有数据点的抽样概率 $P \leftarrow \frac{P}{\|P\|_2}$.

Step 6. 返回概率集合 P.

通过算法 1 得到 n 个样本的索引后,再使用 Nyström 扩展技术对相似矩阵 W 进行逼近,得到 $\hat{W}$,最后,采用谱聚类算法将数据点划分成 k 类.算法流程见算法 2.

算法 2. 基于自适应 Nyström 采样的谱聚类算法(ANS-SC).

输入:数据集 $X=\{x_i|i=1,\ldots,N\}$,采样的数目 n,聚类数目 k;

输出:聚类产生的 k 个簇.

Step 1. 计算 X 中成对数据点的相似性,构造相似度矩阵 $W\in\mathbb{R}^{N\times N}$,其中每个元素 w_{ij} 可以用高斯核函数来表示,即:$w_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)$.

Step 2. 利用算法 1 进行采样,得到 n 个样本的索引集合,然后从 W 中选出相应的元素,组成抽样点间的相似度矩阵 $A\in\mathbb{R}^{n\times n}$,以及抽样点与剩余点的相似度矩阵 $B\in\mathbb{R}^{(N-n)\times n}$.

Step 3. 在矩阵 A 和 B 的基础上,根据公式(5)计算节点的度 $\hat{\boldsymbol{d}}$,然后根据公式(6)、公式(7)分别对矩阵 A 和 B 作归一化处理.

Step 4. 利用归一化的 A,B,计算矩阵 Q:$Q=A+A^{-1/2}BB^TA^{-1/2}$.

Step 5. 将矩阵 Q 对角化:$Q=U_Q\Lambda_QU_Q^T$,得到矩阵 U_Q 和 Λ_Q.

Step 6. 将 U_Q,Λ_Q 代入公式(4),求 $\hat{W}$ 的正交特征向量 V.

Step 7. 从 V 中选出 $\hat{W}$ 前 k 个最大特征值对应的特征向量 $v_1,\ldots,v_k$,形成矩阵 V_k:$V_k=[v_1\vdots\cdots\vdots v_k]\in\mathbb{R}^{N\times k}$.

Step 8. 将矩阵 V_k 的每一行都规范化成单位向量,得到矩阵 Y,其每个元素 $y_{ij}=v_{ij}\Big/\sqrt{\sum_{j=1}^{k}v_{ij}^2}$.矩阵 Y 形成了一个低维嵌入空间$\mathbb{R}^{N\times k}$,其第 i 行与原始数据集的点 x_i 对应.

Step 9. 利用 k-means 算法对矩阵 Y 的行向量进行聚类,若第 i 行被分到第 j 类中,就将原数据点 x_i 归到第 j 类中,这样,将所有数据点划分成 k 个类簇.

算法 1 抽样得到的行的索引都保存在集合 S 中,每次迭代时,在已有采样点的基础上使用一组新的正交向量进行扩展,产生新的样本点 S_i.所选取的每一行的残差二范数$\|E^{(i)}\|^2$以及总的$\|E\|_F^2$,可以通过相似矩阵 W 减去它在 $span(S)$上的正交投影$\pi_S(W)$计算得到.令 M 表示相似矩阵 W 中非零点的数目,N 表示数据集中数据点的总数,n 表示采集的样本点的总数,s 表示一次采样选取的样本数目.每次迭代时,在正交向量上投影所需的时间为 $O(Ms)$.在第 i 次迭代中,由于$\mathbb{R}^N$空间中对 s 个向量进行 Gram-Schmidt 正交化时,一个正交基的大小最多为 $s(i+1)$,所以计算正交基需要的时间为 $O(Ns^2i)$.那么,第 i 次迭代所需的时间为 $O(Ms+Ns^2i)$;t 次迭代,总的时间为 $O(Mst+Ns^2t^2)$.又因为 $t=n/s$,所以算法 1 的时间复杂度为 $O(Mn+Nn^2)$.

算法 2 中输入的数据集一共包含 N 个数据点,Step 1 利用高斯核函数计算这 N 个数据点的相似性值,构造相似矩阵,时间复杂度为 $O(N^2)$;Step 2 通过算法 1 采样,得到 n 个样本的索引,组成矩阵 A 和矩阵 B,时间复杂度为 $O(Mn+Nn^2)$;Step 3 根据公式(5)计算节点度所需的时间为 $O(n(N-n))$,归一化矩阵 A 和 B 需要的时间分别为 $O(n^2)$和 $O(n(N-n))$,因为 $n<<N-n$,所以 Step 3 的时间复杂度为 $O(n(N-n))$;Step 4~Step 6 利用 Nyström 扩展技术求解近似的正交特征向量,时间复杂度为 $O(n^3)$;Step 7 寻找前 k 个特征向量,构造矩阵 V_k 需要的时间为 $O(k)$;Step 8 对 V_k 归一化的时间复杂度为 $O(kN)$;Step 9 使用 k-means 算法聚类的时间复杂度为 $O(kNt)$,其中,t 是迭代的次数.因此,算法 2 的时间复杂度为 $O(N^2)+O(Mn+Nn^2)+O(kNt)$.

5 实验分析

为了对本文提出的 ANS-SC 算法的有效性进行验证,从 UCI 机器学习数据库中选取了 7 个不同大小的数据集,它们的数据特征见表 1.

Table 1 UCI datasets used in the experiments

表 1 实验中使用的 UCI 数据集

数据集	实例点个数	属性数	类数
ImageSeg	2 100	19	7
Musk	6 598	166	2
penDigits	10 992	16	10
mGamma	19 020	10	2
Connect-4	67 557	42	3
USCI	285 779	37	2
Poker Hand	1 000 000	10	3

对于 USCI(US Census Income)数据集,我们去除了包含缺失数据的实例和属性,剩下一共 285 779 个数据点,含有 37 个属性.Poker Hand 数据集非常不均匀,所以我们把小的类聚到一起,而大的类保持不动.这样,最终得到 3 个类,其数据点的数目分别占总实例数的 50.12%,42.25%和 7.63%.同时,我们对 Connect-4 和 USCI 数据集做了归一化处理,使得所有属性的平均值为 0,而标准差为 1.为了确定谱聚类中高斯核参数σ的值,本文采用了交叉验证的方法,搜索范围是[0,200],搜索步长设为 0.1.

实验中,评估聚类性能的指标主要有两个:算法运行时间和聚类准确率.聚类准确率关心的是聚类算法得到的类标签是否正确,需要统计数据集中每个实例的类归属与真实的类标签相符的比例.经过聚类,类的序号往往会重新排列,为了找到聚类结果与原数据集中类划分的对应关系,令 $z=\{1,\ldots,k\}$表示类标签的集合,$\theta(\cdot)$表示真实的类标签,$f(\cdot)$表示聚类算法赋予每个数据点的类标签,那么聚类准确率β可以定义为

$$\beta(f)=\max_{\tau\in\Pi_z}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{\tau(f(x_i))=\theta(x_i)\}\right\} \tag{19}$$

其中,$\mathbb{I}$是指示函数,Π_z是 z 上所有排列的集合.

使用表 1 中的数据集,我们将 ANS-SC 算法与另外 3 种基于不同采样技术的谱聚类算法进行了对比,分别是基于随机抽样的谱聚类算法(RS-SC)[9]、基于加权抽样的谱聚类算法(WS-SC)[19]、基于方差增量抽样的谱聚类算法(IS-SC)[20].实验的计算机环境为:处理器 Pentium(R) Dual-Core E5300 2.60GHz,操作系统 Windows 8,编程环境 MATLAB 2013a.为了客观地对比算法的各项聚类指标,实验中分别将上述 4 种算法执行 20 次,计算其平均聚类准确率和聚类时间.在不同的采样比例下,4 种算法在 UCI 数据集上的聚类准确率以及它们在置信水平为 0.05 下的配对 t 检验的比较结果见表 2.

表 2 中的 P-value 表示在 0.05 的置信水平下,配对 t 检验计算出的 P 值.当 P 值<0.05 时,说明该算法的聚类准确率与 ANS-SC 算法有显著差异;当 P 值>0.05 时,说明该算法的聚类准确率与 ANS-SC 算法的差异不显著.从表 2 中可以看出:随着采集的样本数的增多,各种算法的聚类准确率也不断提高.RS-SC 算法由于采用随机抽样,其聚类性能不太稳定,波动较大,而且聚类准确率也比较低.WS-SC 算法利用了未抽样点的信息,通过最小化舒尔补来降低逼近误差,其聚类表现优于 RS-SC 算法.但是对于大多数数据集,WS-SC 算法的聚类效果不如 IS-SC 算法和 ANS-SC 算法.IS-SC 算法根据特征向量的可聚性,分析舒尔补中相邻数据点之间的关系,然后利用基于方差的启发式策略进行增量抽样,得到所有样本点.在 ImageSeg,Musk,penDigits,mGamma 数据集上,多数情况下,IS-SC 算法的 P 值>0.05,其聚类准确率与 ANS-SC 算法很接近,说明当数据集规模较小时,ANS-SC 算法的聚类表现与 IS-SC 算法差不多.但是在处理大规模的数据集,如 Connect-4,USCI,Poker Hand 时,IS-SC 算法的 P 值<0.05,其聚类准确率明显不如 ANS-SC 算法,说明 ANS-SC 算法的聚类性能更好.可见,ANS-SC 算法使用自适应的采样方法,能够根据数据集自身的结构特征找到更具代表性的样本,使得 Nyström 逼近得到的特征向量更接近真实的特征空间,从而提高聚类的准确率.这 4 种算法在不同数据集上聚类所花费的时间见表 3.

Table 2 Clustering accuracy of algorithms on different datasets (%)

表 2 算法在不同数据集上的聚类准确率(%)

采样比例 n/N (%)	数据集	算法						
		RS-SC		WS-SC		IS-SC		ANS-SC
		准确率	*P*-value	准确率	*P*-value	准确率	*P*-value	准确率
5	ImageSeg	38.64 (±3.35)	0.013	42.35 (±1.67)	0.025	**47.66** (±1.52)	0.872	47.53 (±0.81)
	Musk	53.62 (±3.82)	0.028	54.71 (±1.26)	0.037	59.14 (±1.93)	**0**.648	**60.22** (±1.12)
	penDigits	45.46 (±4.93)	0.046	47.53 (±2.42)	0.051	**51.21** (±1.35)	0.032	49.73 (±0.94)
	mGamma	48.73 (±3.61)	0.019	51.22 (±1.58)	0.024	55.19 (±2.16)	0.739	**56.13** (±1.52)
	Connect-4	41.45 (±4.53)	0.034	**47.61** (±2.74)	0.561	45.92 (±1.84)	0.856	46.17 (±1.48)
	USCI	57.22 (±3.47)	0.021	59.43 (±1.39)	0.036	58.61 (±2.43)	0.027	**64.36** (±0.84)
	Poker Hand	29.87 (±3.86)	0.038	30.65 (±1.52)	0.042	32.52 (±1.28)	0.045	**35.67** (±0.75)
10	ImageSeg	45.12 (±3.41)	0.007	50.17 (±1.84)	0.023	55.49 (±1.32)	0.653	**56.82** (±1.24)
	Musk	67.43 (±4.75)	0.018	69.86 (±2.01)	0.031	**74.63** (±2.61)	0.834	74.15 (±0.63)
	penDigits	56.74 (±3.49)	0.009	58.37 (±1.75)	0.015	65.24 (±1.87)	0.537	**66.37** (±0.74)
	mGamma	63.82 (±3.03)	0.026	**70.31** (±2.24)	0.268	69.55 (±1.56)	0.745	68.19 (±1.33)
	Connect-4	55.26 (±3.78)	0.003	58.38 (±2.18)	0.017	61.14 (±2.93)	0.022	**65.59** (±1.72)
	USCI	72.54 (±4.34)	0.025	75.64 (±1.56)	0.036	75.52 (±1.73)	0.041	**78.48** (±1.35)
	Poker Hand	38.41 (±4.73)	0.004	43.27 (±2.31)	0.029	44.61 (±1.48)	0.038	**49.87** (±0.64)
20	ImageSeg	**75.21** (±3.32)	0.043	70.43 (±2.17)	0.035	72.65 (±1.22)	0.452	73.72 (±0.53)
	Musk	85.67 (±3.98)	0.034	86.29 (±2.05)	0.039	88.21 (±2.56)	0.891	**88.44** (±0.42)
	penDigits	71.53 (±3.41)	0.016	74.93 (±1.51)	0.025	79.75 (±2.29)	0.623	**80.43** (±0.69)
	mGamma	82.44 (±3.65)	0.028	85.34 (±1.68)	0.037	86.47 (±1.78)	0.239	**87.62** (±1.35)
	Connect-4	72.83 (±3.74)	0.005	73.23 (±2.23)	0.014	75.31 (±1.24)	0.023	**78.49** (±1.21)
	USCI	86.72 (±4.13)	0.039	86.88 (±2.74)	0.047	87.35 (±1.58)	0.305	**88.76** (±0.74)
	Poker Hand	49.26 (±3.62)	0.015	50.75 (±2.43)	0.032	53.67 (±1.43)	0.041	**56.84** (±0.83)

Table 3 Clustering time of algorithms on different datasets (s)

表 3 算法在不同数据集上的聚类时间(s)

采样比例 n/N (%)	数据集	算法			
		RS-SC	WS-SC	IS-SC	ANS-SC
5	ImageSeg	0.028	0.563	0.218	0.237
	Musk	0.231	2.034	1.581	1.562
	penDigits	1.046	15.716	9.673	8.741
	mGamma	2.273	25.875	15.425	17.853
	Connect-4	9.165	87.282	59.364	64.235
	USCI	33.251	258.663	181.527	176.338
	Poker Hand	60.547	493.427	320.108	321.176
10	ImageSeg	0.077	1.842	0.441	0.573
	Musk	0.675	5.146	2.237	3.256
	penDigits	3.204	38.749	17.279	21.541
	mGamma	4.296	61.715	38.306	42.304
	Connect-4	13.238	93.924	73.502	76.478
	USCI	45.783	371.852	265.243	284.942
	Poker Hand	73.685	864.217	532.864	577.481
20	ImageSeg	0.263	4.365	1.297	1.963
	Musk	0.814	14.643	6.732	9.805
	penDigits	8.461	77.916	42.961	54.237
	mGamma	10.502	134.285	78.239	95.816
	Connect-4	18.032	213.328	149.343	178.543
	USCI	66.394	587.572	308.642	325.776
	Poker Hand	117.673	1293.631	813.524	876.312

由表 3 可知:样本点的数目越多,算法聚类所需要的时间也越长.基于随机抽样的 RS-SC 算法在各个数据集上运行的时间都是最少的,因为另外 3 种算法使用改进的抽样策略,需要花费额外的时间寻找合适的样本点.其中,WS-SC 算法耗时最长,主要由于它在每一次采样迭代时,都需要计算一个 n 阶矩阵的行列式,算法复杂度较高.当采样比例为 5%时,ANS-SC 算法的运行时间与 IS-SC 算法差不多.但是当采样比例增加到 10%和 20%时,ANS-SC 算法在几个数据集上聚类的时间都长于 IS-SC 算法,说明自适应 Nyström 采样方法所需的计算量还是

比较大的.如何有效降低其时间复杂度,今后需要进一步研究.

6 结束语

Nyström 矩阵低秩逼近方法可以求解近似的特征向量和特征空间,是处理大数据的一个有力工具.在谱聚类中,利用 Nyström 扩展技术进行近似计算,可以在很大程度上降低时间和空间的开销,以较小的精度损失换取算法效率的大幅提升.使用 Nyström 方法时,抽样策略的选择会对聚类结果的优劣产生重要影响.为了使所选取的少数样本点更好地反映数据集的分布情况,设计了一种自适应的采样方法用于 Nyström 扩展技术.该方法通过多次遍历,根据不同的概率进行采样,每次采样只选择一部分样本点,并更新剩余点的采样概率.如此迭代循环,直到已经选择了足够的样本点为止.本文从理论上证明了采样次数的增加可以有效降低抽样误差.接着,提出一种基于自适应 Nyström 采样的谱聚类算法.在 UCI 机器学习数据库上的实验表明:该算法的聚类效果优于基于随机抽样的谱聚类算法,基于加权抽样的谱聚类算法以及基于方差增量抽样的谱聚类算法,能够较好地处理大规模的数据集.

致谢 在此,我们向对本文的工作给予支持和建议的同行表示感谢.

References:

[1] Sun JG, Liu J, Zhao LY. Clustering algorithms research. Ruan Jian Xue Bao/Journal of Software, 2008,19(1):48−61 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048]

[2] Ding SF, Jia HJ, Zhang LW, Jin FX. Research of semi-supervised spectral clustering algorithm based on pairwise constraints. Neural Computing and Applications, 2014,24(1):211−219. [doi: 10.1007/s00521-012-1207-8]

[3] Chen XL, Deng C. Large scale spectral clustering with landmark-based representation. In: Proc. of the 25th AAAI Conf. on Artificial Intelligence. 2011. 313−318.

[4] Song YQ, Chen WY, Bai HJ, Lin CJ, Chang EY. Parallel spectral clustering. Machine Learning and Knowledge Discovery in Databases, 2008, 5212:374−389. [doi: 10.1007/978-3-540-87481-2_25]

[5] Yan DH, Huang L, Jordan MI. Fast approximate spectral clustering. In: Proc. of the 15th ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD). 2009. 907−916. [doi: 10.1145/1557019.1557118]

[6] Lin F, Cohen WW. Power iteration clustering. In: Proc. of the Int'l Conf. on Machine Learning. 2010. 655−662.

[7] Li M, Kwok JT, Lu BL. Making large-scale Nyström approximation possible. In: Proc. of the Int'l Conf. on Machine Learning. 2010. 631−638.

[8] Williams CKI, Seeger M. Using the Nyström method to speed up kernel machines. In: Proc. of the Advances in Neural Information Processing Systems 13. 2001. 682−688.

[9] Fowlkes C, Belongie S, Chung F, Malik J. Spectral grouping using the Nyström method. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004,26:214−225. [doi: 10.1109/TPAMI.2004.1262185]

[10] Kumar S, Mohri M, Talwalkar A. Ensemhle Nyström method. In: Proc. of the Advances in Neural Information Processing Systems. 2009. 1060−1068.

[11] Zhang K, Tsang IW, Kwok JT. Improved Nyström low-rank approximation and error analysis. In: Proc. of the 25th Int'l Conf. on Machine Learning. 2008. 1232−1239. [doi: 10.1145/1390156.1390311]

[12] Zhang K, Kwok JT. Clustered Nyström method for large scale manifold learning and dimension reduction. IEEE Trans. on Neural Networks, 2010,21(10):1576−1587. [doi: 10.1109/TNN.2010.2064786]

[13] Wang L, Bezdek JC, Leckie C, Kotagirl R. Selective sampling for approximate clustering of very large data sets. Int'l Journal of Intelligent Systems, 2008,23(3):313−331. [doi: 10.1002/int.20268]

[14] Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8): 888−905. [doi: 10.1109/34.868688]

[15] Dhanjal C, Gaudel R, Clemencon S. Efficient eigen-updating for spectral graph clustering. Neurocomputing, 2014,131:440−452. [doi: 10.1016/j.neucom.2013.11.015]

[16] Shi J, Malik J. Motion segmentation and tracking using normalized cuts. In: Proc. of the Int'l Conf. on Computer Vision. 1998. 1154−1160. [doi: 10.1109/ICCV.1998.710861]

[17] Frieze A, Kannan R, Vempala S. Fast Monte-Carlo algorithms for finding low-rank approximations. Journal of the ACM, 2004,51: 1025−1041. [doi: 10.1145/1039488.1039494]

[18] Drineas P, Frieze A, Kannan R, Vempala S, Vinay V. Clustering large graphs via the singular value decomposition. Machine Learning, 2004,56: 9−33. [doi: 10.1023/B:MACH.0000033113.59016.96]

[19] Belabbas M, Patrick JW. Spectral methods in machine learning and new strategies for very large datasets. Proc. of the National Academy of Sciences of the USA, 2009,51(6):369−374. [doi: 10.1073/pnas.0810600105]

[20] Zhang XC, You QZ. Clusterability analysis and incremental sampling for Nyström extension based spectral clustering. In: Proc. of the IEEE 11th Int'l Conf. on Data Mining (ICDM). 2011. 942−951. [doi: 10.1109/ICDM.2011.35]

附中文参考文献:

[1] 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48−61. http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.3724/SP.J.1001.2008.00048]

丁世飞(1963－),男,山东青岛人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能,机器学习,数据挖掘,粒度计算.
E-mail: dingsf@cumt.edu.cn

贾洪杰(1988－),男,博士生,主要研究领域为感知计算,谱聚类,机器学习.
E-mail: jiahongjie@cumt.edu.cn

史忠植(1941－),男,研究员,博士生导师,CCF 高级会员,主要研究领域为智能科学,人工智能,机器学习.
E-mail: shizz@ics.ict.ac.cn