关于二部图谱聚类泛化性的研究

梁伟轩 刘新旺 蓝 龙 祝 恩
(国防科技大学计算机学院 长沙 410073)

摘 要 谱聚类算法是一种重要的聚类算法,能够在多种应用场景中取得理想的聚类效果,但较高的计算复杂度限制了其在大规模数据集上的应用。为了提高计算效率,研究者开发了二部图谱聚类算法。具体来说,此类方法仅选取部分训练集作为锚点集,并利用整个训练集和锚点集构建二部图,再利用该二部图进行近似的谱聚类。然而,这类方法存在以下三个没有被充分研究的问题:一是二部图谱聚类算法是否具备泛化性;二是如何快速获取训练集外顶点的低维嵌入;三是如何选择锚点数规模,使算法达到统计精度和计算开销的最佳平衡。针对上述三个问题,本文先是建立了谱聚类泛化分析的框架,并根据谱聚类的一致性,推导了标准NCut算法的泛化风险上界和额外风险上界。接着,本文分析了针对标准NCut的一种近似算法的泛化性,即基于Nyström方法的二部图谱聚类算法。根据所得到的二部图谱聚类的泛化理论,本文提出了一种能够快速获取训练集外顶点低维嵌入的算法。此外,本文还通过上述理论提出了一种锚点数选择的策略,即锚点数为$\Theta(\sqrt{n})$时,算法达到统计精度与计算效率的最佳平衡。最后,本文在基准数据集上验证了所提出算法的有效性和理论结果的正确性。

关键词 谱聚类;二部图;泛化分析;聚类风险
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.01065

On the Generalization of Spectral Clustering on Bipartite Graph

LIANG Wei-Xuan LIU Xin-Wang LAN Long ZHU En
(College of Computer Science and Technology, National University of Defense Technology, Changsha 410073)

Abstract Spectral clustering is an important clustering algorithm that achieves desirable clustering performance in various application scenarios. However, its high computational complexity limits its applicability to large-scale datasets. To improve computational efficiency, researchers have developed spectral clustering algorithms on bipartite graphs. Specifically, these methods select only a subset of the training set as an anchor set and construct a bipartite graph using the whole training set and the anchor set, performing approximate spectral clustering on the bipartite graph. However, there are three issues with such methods that have not been thoroughly studied: (1) whether the spectral clustering algorithm on bipartite graph possesses generalization ability; (2) how to efficiently obtain low-dimensional embeddings for the out-of-sample points; (3) how to determine the scale of anchor number to obtain the optimal trade-off between statistical accuracy and computational efficiency. To address the above three issues, this paper first establishes a framework for analyzing the generalization of spectral clustering and, based on the consistency of spectral clustering, derives the upper bounds of both the generalization risk and the excess risk of the standard NCut algorithm. Then, this paper also analyzes the generalization ability of an approximation algorithm for the standard NCut, i.e., the Nyström-based spectral

收稿日期:2024-12-14;在线发布日期:2025-03-05。本课题得到国家自然科学基金面上项目(No. 62276271)、国家自然科学基金杰出青年科学基金项目(No. 62325604)资助。**梁伟轩**,博士,博士后,主要研究领域为多视图聚类、核学习和学习理论。E-mail:weixuanliang@nudt.edu.cn。**刘新旺(通信作者)**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为多视图聚类、核学习。E-mail:xinwangliu@nudt.edu.cn。**蓝 龙**,博士,副研究员,主要研究领域为目标检测、目标重识别和聚类算法。**祝 恩**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为聚类算法、异常检测和模式识别。

clustering algorithm on bipartite graph. Based on the derived generalization theory for spectral clustering on bipartite graph, this paper proposes an algorithm that can obtain low-dimensional embeddings for out-of-sample points. Furthermore, a theoretical strategy for selecting the number of anchor points is proposed, revealing that when the number of anchor points is $\Theta(\sqrt{n})$, the algorithm achieves the optimal trade-off between statistical accuracy and computational efficiency. Finally, the proposed algorithm's effectiveness and the correctness of the theoretical results are validated on benchmark datasets.

Keywords spectral clustering; bipartite graph; generalization analysis; clustering risk

1 引 言

谱聚类[1]是一种基础的聚类算法,其基于数据样本构建图矩阵,再利用特征分解获取顶点的低维嵌入,最后利用顶点的低维嵌入得到最后的聚类结果。因为谱聚类能够有效地处理线性不可分的数据,在诸多学习任务中都有成功的应用,取得了非常好的效果,例如:图像分割[2-3]、目标检测[4-5]、社区检测[6-7]、基因表达分析[8-9]、语音分割[10-11]和多视图聚类[12-16]等等。然而,谱聚类算法的计算复杂度高达$O(n^3)$,其中n为图矩阵的顶点数。这限制了谱聚类算法在大规模数据集上的应用。

针对高计算复杂度的问题,研究者们发现谱聚类中存在非常高的计算冗余,即不需要构建整个图矩阵即能获得类似的聚类结果[17-20]。这些方法从顶点集中选取部分顶点作为锚点,构建顶点与锚点之间的二部图矩阵,再利用近似算法,在二部图上进行近似的谱聚类算法。因此,在本文中,这类方法被统称为二部图谱聚类算法。文献[17]和文献[18]在构建二部图矩阵以后,利用邻居信息构建了顶点与锚点间的邻接矩阵,最终通过近似全图矩阵的邻接矩阵得到最终样本低维嵌入。而文献[19]则是利用二部图矩阵直接近似全图的 Laplacian 矩阵来获得低维嵌入。还有一类方法,则是基于 Nyström 方法[20]。Nyström 方法本是一种用于积分方程求数值解的算法[21],后常用于加速核学习算法[22-23]。基于 Nyström 方法的二部图谱聚类算法通过近似全图矩阵以及相应的度矩阵,以达到近似标准谱聚类算法的效果[20]。此类方法中,算法统计精度会随锚点数减少而降低,但计算开销也会相应地降低,反之亦然。这些方法自提出以后便在各个领域取得了重要应用[24-27],有效实现了算法加速的同时,也保证了一定的聚类效果。

上述方法虽然有效地对谱聚类进行了加速,使计算复杂度从$O(n^3)$降低至$O(n)$,然而这些方法还存在以下三个亟待解决的问题:一是缺乏必要的理论保证,没有对算法的泛化性进行分析;二是难以高效地获取训练集外顶点的低维嵌入;三是如何选择锚点数规模,使算法达到统计精度和计算开销的最佳平衡。本文试图对谱聚类算法中的 NCut 算法及其二部图近似算法的泛化性展开研究,解决上述三个问题。

泛化分析是统计学习理论中的重要领域,其旨在衡量算法的经验解在整个空间上的表现。因为缺少标签,聚类算法的泛化分析一直是一个非常困难的课题。针对k均值和核k均值聚类算法,文献[28]基于假设空间一致收敛的原则,利用 Rademacher 复杂度得到了$O(k/\sqrt{n})$的额外风险界,其中k为聚类簇数。文献[29]则将上述风险界改进至$\widetilde{O}(\sqrt{k/n})$①,并通过下界的匹配证明了所推导的上界的最优性。对于投影聚类,文献[30]给出了此类方法的泛化界。关于谱聚类算法的泛化分析,文献[31]基于对 U-统计量的研究,将点对累加转化为单层累加的形式,并利用 Rademacher 复杂度给出了谱聚类的泛化界。

本文先是建立了谱聚类算法的泛化理论框架,定义了谱聚类的损失函数、经验损失和期望损失。根据上述定义,还给出了泛化风险和额外风险的定义,并利用谱聚类的一致性[32],推导出其泛化风险和额外风险均有上界为$\widetilde{O}(\sqrt{k/n})$。本文证明所使用的方法,仅使用了基础的 Hoeffding 不等式[33]以及文献[32]所建立的谱聚类算法一致性结论,得到了谱聚类的泛化界。本文所用的方法,相比于文献[31]的方法,更为简单和直观。

接着,本文对基于 Nyström 方法的二部图谱聚

① $\widetilde{O}(\cdot)$隐藏了对数项,指隐藏了关于$\log(kn/\delta)$的多项式项。

类算法进行了泛化分析。本文先是给出了二部图谱聚类对应的聚类指示函数,并依照前面所建立的理论框架,对二部图谱聚类的泛化风险界进行了推导。依照泛化风险界的结论,本文利用经验 Laplacian 算子的经验特征函数,提出了一种能够快速计算训练集外顶点低维嵌入的算法。此外,本文还推导了相应的额外风险界,并据此给出了使统计精度和计算效率达到最佳平衡的锚点规模。最后,本文还在数个基准数据集上进行了综合的实验。实验首先测试了不同锚点数与基于 Nyström 方法的二部图谱聚类算法对于原算法的近似程度之间的关系,验证了所推导的额外风险界的正确性。此外,本文还验证了所提出的快速获取训练集外顶点低维嵌入算法的有效性。

本文的主要贡献总结如下:

(1)通过谱聚类一致性的结论,以一种更简洁直观的方式推导出标准 NCut 谱聚类算法的泛化风险界和额外风险界均为 $\widetilde{O}(k/\sqrt{n})$(定理 1)。

(2)给出了基于 Nyström 方法的二部图谱聚类算法的聚类指示函数,推导出该方法的泛化聚类风险界为 $\widetilde{O}(k/\sqrt{n})$(定理 2),并据此提出一种快速获取训练集外顶点低维嵌入的算法。

(3)给出了基于 Nyström 方法的二部图谱聚类算法的额外聚类风险界与锚点数之间的关系(定理 3),并且提供了一种锚点数选择方案,即锚点数为 $\Theta(\sqrt{n})$ 时,能够使该方法达到最佳统计精度与计算效率的平衡。

2　相关工作

符号系统和基本假设. 本文在介绍相关工作之前,先对本文相关的符号系统和假设进行介绍。本文使用小写的粗体字母表示向量,大写的粗体字母表示矩阵;用 $\|\cdot\|$ 表示范数:若 $\boldsymbol{a}$ 是向量,$\|\boldsymbol{a}\|$ 表示向量 2 范数,$\|\boldsymbol{a}\|_\infty$ 表示无穷大范数,即向量 $\boldsymbol{a}$ 所有分量绝对值最大者;若 $\boldsymbol{A}$ 是矩阵,$\|\boldsymbol{A}\|$ 为矩阵谱范数,$\|\boldsymbol{A}\|_F$ 为矩阵的 Frobenius 范数。用 $\rho_n(\cdot)$ 和 $\rho(\cdot)$ 分别表示顶点集的经验分布函数和真实分布函数。用 χ 表示顶点集所属的空间。若矩阵 $\boldsymbol{A}$ 可逆,$\boldsymbol{A}^{-1}$ 表示矩阵的逆;否则 $\boldsymbol{A}^{-1}$ 表示矩阵的伪逆。设 f,g 分别为两个函数,$f(n)=O(g(n))$ 表示 $f(n)<cg(n)$,$f(n)=O(g(n))$ 表示 $f(n)\leqslant cg(n)$,$f(n)=\Theta(g(n))$ 表示 $c_1g(n)\leqslant f(n)\leqslant c_2g(n)$,其中 c、c_1、c_2 为正常数。表 1 列出了符号的基本说明。主要假设参考文献[32],即存在正常数 l,b,任意两个顶点 $\boldsymbol{x},\boldsymbol{y}\in\chi$ 的相似度函数满足 $l\leqslant W(\boldsymbol{x},\boldsymbol{y})\leqslant b$,且文中所涉及的矩阵均只有单重特征值。

表 1　符号的基本说明

符号	含义
小写粗体字母	向量
大写粗体字母	矩阵
$\rho_n(\cdot)$	经验分布函数
$\rho(\cdot)$	真实分布函数
$\|\boldsymbol{a}\|$	向量 $\boldsymbol{a}$ 的 2 范数
$\|\boldsymbol{a}\|_\infty$	向量 $\boldsymbol{a}$ 的无穷大范数
$\|\boldsymbol{A}\|$	矩阵 $\boldsymbol{A}$ 的谱范数
$\|\boldsymbol{A}\|_F$	矩阵 $\boldsymbol{A}$ 的 Frobenius 范数
$o(\cdot)$	非渐近紧确上界
$O(\cdot)$	渐近上界
$\widetilde{O}(\cdot)$	隐藏了对数项的渐近上界
$\Theta(\cdot)$	渐近紧确界
$W(\cdot,\cdot)$	相似度函数
χ	顶点空间

2.1　谱聚类

谱聚类[1]主要分为两种,分别是 Ratio Cut(本文简写为 RCut)和 Normalized Cut(本文简写为 NCut)。根据研究谱聚类一致性的文献所述,相比于 RCut,NCut 具备更好的统计性质[32],因此本文主要以 NCut 为研究对象。NCut(松弛形式)的目标式如下

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{k}W(\boldsymbol{x}_i,\boldsymbol{x}_j)(h_{it}-h_{jt})^2 \tag{1}$$

其中设 $\boldsymbol{H}\in\mathcal{R}^{n\times k}$,$h_{it}$ 为位于矩阵 $\boldsymbol{H}$ 第 i 行第 t 列的元素;$W(\cdot,\cdot)$ 为相似度函数,本文设 $W(\cdot,\cdot)$ 为高斯核函数,即 $W(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\sigma^2)$,$\sigma$ 为核函数宽度参数。设 $\boldsymbol{D}$ 为对角阵,其对角元素为顶点的度 $d_{jj}=\frac{1}{n}\sum_{i=1}^{n}W(\boldsymbol{x}_i,\boldsymbol{x}_j)$,式 (1) 满足约束 $n\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{H}=\boldsymbol{I}_k$。令 $\boldsymbol{F}=\sqrt{n}\boldsymbol{D}^{1/2}\boldsymbol{H}$,式 (1) 可转换为如下矩阵形式:

$$\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{F}) \tag{2}$$

其中 $\boldsymbol{L}\in\mathcal{R}^{n\times n}$ 为 Laplacian 矩阵,满足 $\boldsymbol{L}=\boldsymbol{I}_n-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$,且 $\boldsymbol{F}$ 满足 $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}=\boldsymbol{I}_k$。式 (2) 的最优解为 $\boldsymbol{L}$ 最小的 k 个特征值对应的特征向量 $\boldsymbol{F}$。最终,NCut 得到聚类指示矩阵为 $\boldsymbol{H}=\frac{1}{\sqrt{n}}\boldsymbol{D}^{-1/2}\boldsymbol{F}$,并在 $\boldsymbol{H}$ 上执行标准的 k 均值聚类算法,得到最终的聚类结果。

2.2 二部图谱聚类

通过上一节的介绍可知，NCut 存在一个无法避开的问题，那就是较高的复杂度。由于需要构建一个规模为 $n \times n$ 的 Laplacian 矩阵，所以空间复杂度至少为 $O(n^2)$。后续还需要对 $\boldsymbol{L}$ 进行特征分解，其时间复杂度为 $O(n^3)$。当数据集顶点比较多时，较高的复杂度将使谱聚类难以运行。为了降低复杂度，文献[17-20]采用了二部图来近似全图的策略，再设计能够在二部图上运行的谱聚类算法。

文献[17-18]先随机地从整个顶点集 $\{\boldsymbol{x}_i\}_{i=1}^n$ 中选出 m 个锚点 $\{\boldsymbol{a}_j\}_{j=1}^m$，计算出二部图矩阵 $\boldsymbol{P} \in \mathcal{R}^{n\times m}$，其元素可以表示为 $P_{ij}=W(\boldsymbol{x}_i,\boldsymbol{a}_j)$。接着，文献[17-18]利用 $\boldsymbol{P}$ 构造了邻接矩阵 $\boldsymbol{Z} \in \mathcal{R}^{n\times m}$，其元素为

$$Z_{ij}=\frac{P_{ij}}{\sum_{j\in\mathcal{N}_s(\boldsymbol{x}_i)}P_{ij}} \tag{3}$$

其中 $\mathcal{N}_s(\boldsymbol{x}_i)$ 表示距离顶点 $\boldsymbol{x}_i$ 最近的 s 个锚点的索引，即集合 $\{P_{ij}\}_{j=1}^m$ 中最大的 s 个元素的列角标。整个相似度矩阵可用 $\boldsymbol{Z}\Lambda^{-1}\boldsymbol{Z}^{\mathrm{T}}$ 进行表示，其中 Λ 为对角矩阵，其第 j 个对角元素 $d_{jj}=\sum_{i=1}^n Z_{ij}$。最后可对 $\boldsymbol{Z}\Lambda^{-1/2}$ 进行奇异值分解，得到其前 k 个最大奇异值对应的左奇异向量，再将这些向量作为最后的聚类指示矩阵得到最终的聚类结果。在文献[19]中，二部图矩阵 $\boldsymbol{P}$ 的构造如前所述，但其先计算出由 $\boldsymbol{P}$ 的行和与列和组成的对角矩阵 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$。接着，直接利用 $\boldsymbol{D}_1^{-1/2}\boldsymbol{P}\boldsymbol{D}_2^{-1/2}$ 对二部图 $\boldsymbol{P}$ 进行标准化，然后再求得其最大的 k 个左奇异向量作为聚类指示矩阵。

上述两种方法最多需要存储一个 $n\times m$ 维的矩阵，故它们空间复杂度为 $O(nm)$；时间消耗最多的地方是对一个 $n\times m$ 维矩阵进行奇异值分解，其时间复杂度最多为 $O(nm^2)$。当 $m \ll n$ 时，上述两种方法能有效降低谱聚类的复杂度。然而，在这些方法中，文献[17-18]中的方法并不是对原始的谱聚类进行近似；而文献[19]的方法并没有对原 Laplacian 矩阵进行有效地近似。可以看出，上述两种方法均是基于直观而构造的算法，其获得的聚类指示矩阵与原算法聚类指示矩阵的偏离程度是无法被估计的，因此难以研究其统计意义上的近似效果。为此，本文拟采用基于 Nyström 法二部图谱聚类算法[20]，并研究该方法的统计性质。下面先对 Nyström 法[34]进行一个简要的介绍。

2.3 Nyström 方法

Nyström 方法[34]常用于近似核矩阵，并且被应用于加速核 k 均值聚类算法[35-36]。其方法的具体流程如下，给定一个半正定矩阵 $\boldsymbol{W} \in \mathcal{R}^{n\times n}$，对 $\boldsymbol{W}$ 进行列采样 m 列构造矩阵 $\boldsymbol{P} \in \mathcal{R}^{n\times m}$。同时，设 $\boldsymbol{K} \in \mathcal{R}^{m\times m}$ 为对 $\boldsymbol{P}$ 行采样 m 行构造的矩阵，其中行采样的索引对应上述 m 列的索引。那么，$\boldsymbol{W}$ 可被 $\boldsymbol{P}\boldsymbol{K}^{+}\boldsymbol{P}^{T}$ 近似，文献[34]给出了 $\|\boldsymbol{W}-\boldsymbol{P}\boldsymbol{K}^{+}\boldsymbol{P}^{T}\|_F$ 的误差上界。同时，$\boldsymbol{P}\boldsymbol{K}^{+1/2} \in \mathcal{R}^{n\times m}$ 的行向量可被看作样本的 m 维的表示，作为标准 k 均值聚类算法的输入，用于加速核 k 均值聚类算法。相应地，文献[35]和文献[36]给出了上述近似方法的聚类损失关于标准核 k 均值聚类算法聚类损失的差异上界。Nyström 方法是一种具备良好理论保证，且应用广泛的方法。

3 谱聚类泛化的理论框架及结论

本节先介绍谱聚类泛化分析的理论框架。评判一个聚类算法好坏，除了看训练损失的大小，更重要的是其在全空间上的表现。为此，需要对谱聚类的聚类损失及聚类风险进行定义。为了研究全空间上谱聚类算法的损失，需要先对期望版本的 NCut 的 Laplacian 算子进行定义。附录 1 对算子理论的基础进行了基本介绍。首先，依照关于谱聚类一致性的研究[32]所述，引入如下几个算子的定义。

定义 1. 谱聚类中的经验度算子与期望度算子分别定义如下：

$$\begin{aligned} d_n(\boldsymbol{x})&=\int_X W(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\rho_n(\boldsymbol{y}),\\ d(\boldsymbol{x})&=\int_X W(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\rho(\boldsymbol{y}) \end{aligned} \tag{4}$$

定义 2. NCut 的 Laplacian 算子。设经验的和期望的标准化相似度函数分别为

$$G_n(\boldsymbol{x},\boldsymbol{y})=\frac{W(\boldsymbol{x},\boldsymbol{y})}{\sqrt{d_n(\boldsymbol{x})d_n(\boldsymbol{y})}},G(x,y)=\frac{W(\boldsymbol{x},\boldsymbol{y})}{\sqrt{d(\boldsymbol{x})d(\boldsymbol{y})}} \tag{5}$$

那么 NCut 中的经验 Laplacian 算子 L_n 和期望 Laplacian 算子 L 有如下定义：

$$\begin{aligned} L_nf(\boldsymbol{x})&=f(\boldsymbol{x})-\int_X G_n(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{y})\mathrm{d}\rho_n(\boldsymbol{y}),\\ Lf(\boldsymbol{x})&=f(\boldsymbol{x})-\int_X G(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{y})\mathrm{d}\rho(\boldsymbol{y}) \end{aligned} \tag{6}$$

下面的命题给出了 NCut 中 Laplacian 矩阵 $\frac{1}{n}\boldsymbol{L}$ 和经验 Laplacian 算子 L_n 的谱之间的关系。

命题 1.(文献[32]的命题 9)　$\frac{1}{n}\boldsymbol{L}$ 和 L_n 拥有相同的非 0 特征值。设第 t 个特征值为 $\hat{\lambda}_t$,那么 $\frac{1}{n}\boldsymbol{L}$ 第 t 个($t\in[n]$,按对应特征值大小升序排列)的特征向量 $\boldsymbol{f}_t=[f_{1t},\cdots,f_{nt}]^{\mathrm{T}}$ 与 L_n 的第 t 个特征函数 $\hat{f}_t(\cdot)$ 有如下关系:

$$\forall j\in[n],f_{jt}=\hat{f}_t(\boldsymbol{x}_j)/\sqrt{n},$$
$$\forall \boldsymbol{x}\in X,\hat{f}_t(\boldsymbol{x})=\frac{1/n\sum_{j=1}^{n}G_n(\boldsymbol{x},\boldsymbol{x}_j)\hat{f}_t(\boldsymbol{x}_j)}{1-\hat{\lambda}_t} \tag{7}$$

设期望的 Laplacian 算子 L 的特征值为 $\{\lambda_t\}_{t=1}^{+\infty}$。根据文献[32]中的定理 15(参见附录 1),可知随着 $n\to+\infty$时,有 $\hat{\lambda}_t\xrightarrow{a.s.}\lambda_t$。同时,由文献[32]中的定理 16 及例 1(参见附录 1),可知存在一组无穷序列 $\{a_n\}_{n=1}^{+\infty}$,其元素 $a_n=1$ 或 -1,使 $a_n\hat{f}_t\xrightarrow{a.s.}f_t^*$。设 L 中的特征对为 $\{(\lambda_t,f_t^*)\}_{t=1}^{k}$。由聚类指示矩阵 $\boldsymbol{H}$ 与 Laplacian 矩阵 $\boldsymbol{L}$ 的特征向量 $\boldsymbol{F}$ 的关系可知,相应的聚类指示函数为

$$\hat{H}=\{\hat{h}_t\}_{t=1}^{k},\text{其中}\hat{h}_t(\boldsymbol{x})=\frac{\hat{f}_t(\boldsymbol{x})}{\sqrt{d_n(\boldsymbol{x})}} \tag{8}$$

类似地,可知期望版本 NCut 的聚类指示函数为

$$H^*=\{h_t^*\}_{t=1}^{k},\text{其中}h_t^*(\boldsymbol{x})=\frac{f_t^*(\boldsymbol{x})}{\sqrt{d(\boldsymbol{x})}} \tag{9}$$

定义 3. 对于空间 X 中任意两个顶点 $\boldsymbol{x}$ 和 $\boldsymbol{y}$,谱聚类的成对损失函数由下式定义

$$l(\boldsymbol{x},\boldsymbol{y},H)=\sum_{t=1}^{k}W(\boldsymbol{x},\boldsymbol{y})(h_t(\boldsymbol{x})-h_t(\boldsymbol{y}))^2 \tag{10}$$

其中 $H=\{h_t(\cdot)\}_{t=1}^{k}$ 为某组聚类指示函数。据此,定义 NCut 的经验聚类风险 $\hat{R}(H)$ 为

$$\frac{1}{2n^2}\sum_{i,j=1}^{n}\sum_{t=1}^{k}W(\boldsymbol{x}_i,\boldsymbol{x}_j)(h_t(\boldsymbol{x}_i)-h_t(\boldsymbol{x}_j))^2 \tag{11}$$

相应地,期望聚类风险为

$$R(H)=\frac{1}{2}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\sum_{t=1}^{k}W(\boldsymbol{x},\boldsymbol{y})(h_t(\boldsymbol{x})-h_t(\boldsymbol{y}))^2\right] \tag{12}$$

注意到,在经验版本的 NCut 中,$\boldsymbol{D}^{1/2}\boldsymbol{H}$ 中的列向量是单位正交的。类似地,期望版本的 NCut 就是要找到一组在空间 χ 上聚类指示函数 $\{h_t(\cdot)\}_{t=1}^{k}$ 使得 $R(H)$ 最小,且 $\{\sqrt{d(\cdot)}h_t(\cdot)\}_{t=1}^{k}$ 也满足在空间 X 上的单位正交约束。因此,式(9)所述 H^* 恰为满足上述约束,使得 $R(H)$ 最小的解。据此,可以定义谱聚类的泛化风险和额外风险。

定义 4. 泛化风险和额外风险。根据谱聚类的经验聚类风险及期望聚类风险的定义,可以定义如下 NCut 的泛化风险

$$R(\hat{H})-\hat{R}(\hat{H}) \tag{13}$$

其中 $\hat{H}$ 为某组聚类指示函数,一般通过经验算法得到。相应地,$\hat{H}$ 的额外风险为

$$R(\hat{H})-R(H^*) \tag{14}$$

上述泛化风险是衡量学习算法得到的参数在整个顶点空间上的表现是否能够接近在训练集上的表现。泛化风险越小,就说明算法的泛化性能越好。而额外风险是用于衡量学习算法得到的参数和整个顶点空间中最优参数的差异。额外风险越小,就说明经验参数能够更好地近似整个顶点空间中的最优参数。本文首先给出如下 NCut 的泛化风险界和额外风险界。

定理 1. 由经验 NCut 算法得到的聚类指示函数 $\hat{H}$(如式(8)所示)的泛化风险和额外风险均有上界 $\widetilde{O}(k/\sqrt{n})$。

注释. 本文利用 NCut 谱聚类中特征函数具备一致性的性质,首先给出了 NCut 算法的泛化风险和额外风险界。具体的证明过程请见附录 2。这个定理所述的泛化风险界说明,随着 n 变大,经验聚类指示函数的泛化性会越来越好。与此同时,定理所述的额外风险界说明,经验聚类指示函数会随着 n 变大,更加接近最优的聚类指示函数。该定理从理论上证明了 NCut 算法具备良好的泛化性。该定理的证明参见附录 2。

4　二部图谱聚类算法及泛化分析

本节的第一部分先对基于 Nyström 方法的二部图谱聚类算法进行介绍,并给出其相应的聚类指示函数。接着,本节推导了该方法的泛化聚类风险上界,并据此给出了一种能够快速获取训练集外顶点低维嵌入的方法。本节还推导了该方法的额外聚类风险上界,并给出了一种锚点规模选择的策略。

4.1　基于 Nyström 的二部图谱聚类算法

Nyström 方法常用于加速核学习算法,并且具备统计理论保证,故本文选用该方法对二部图进行处理。NCut 中,要对 $\boldsymbol{L}=\boldsymbol{I}_n-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ 进行特征分解,取 $\boldsymbol{L}$ 最小的 k 个特征值对应的特征向量,即

$\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ 最大的 k 个特征值对应的特征向量 $\boldsymbol{F}$。下面利用二部图对 $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ 进行近似。

根据 2.2 节的描述，假设已经随机选取了 m 个锚点，并且利用这 m 个锚点与整个顶点集构建了二部图矩阵 $\boldsymbol{P}\in\mathcal{R}^{n\times m}$。与此同时，构建锚点与锚点之间的图矩阵 $\boldsymbol{K}\in\mathcal{R}^{m\times m}$，即按照锚点索引对 $\boldsymbol{P}$ 行采样得到的矩阵。对锚点间的图矩阵特征分解，即 $\boldsymbol{K}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}$。接着，可用 $\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}$ 近似全部顶点集的图矩阵 $\boldsymbol{W}$，顶点集的度可由 $\widetilde{\boldsymbol{D}}=diag(\frac{\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{1}_n}{n})$ 进行近似。此处为了降低存储复杂度，并不直接计算出上述对角矩阵，而是仅利用一个 n 维向量存储其对角元素。最后，对 $\widetilde{\boldsymbol{D}}^{-1/2}\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$ 进行特征分解，取其最大的 k 个左奇异值对应的奇异向量 $\widetilde{\boldsymbol{F}}$ 作为 $\boldsymbol{F}$ 的近似。再令 $\widetilde{\boldsymbol{H}}=\widetilde{\boldsymbol{D}}^{-1/2}\widetilde{\boldsymbol{F}}$ 为聚类指示矩阵的近似，作为标准 k 均值聚类算法的输入，获取最终的聚类结果。其算法流程如算法 1 所示。

算法 1. 基于 Nyström 的二部图谱聚类算法

输入：顶点集 $\{\boldsymbol{x}_i\}_{i=1}^n$，锚点集 $\{\boldsymbol{a}_j\}_{j=1}^m$，聚类簇数 k，高斯核宽度参数 σ

输出：聚类结果

1. 利用高斯核函数计算出顶点集与锚点集的二部图矩阵 $\boldsymbol{P}$，以及锚点集之间的图矩阵 $\boldsymbol{K}$
2. 对 $\boldsymbol{K}$ 进行特征分解，即 $\boldsymbol{K}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}$。通过
$$\widetilde{\boldsymbol{D}}=diag(\frac{\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{1}_n}{n})$$
计算顶点集近似的度矩阵，设其第 i 个元素为 $\widetilde{d}_{ii}$
3. 计算矩阵 $\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$
4. FOR $\forall i=1,\cdots,n$ DO
5. 计算 $\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-\frac{1}{2}}(i,:)=\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-\frac{1}{2}}(i,:)/\sqrt{\widetilde{d}_{ii}}$
6. END FOR
7. 对 $\widetilde{\boldsymbol{D}}^{-1/2}\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$ 进行奇异值分解，得到其前 k 个最大奇异值对应的左奇异向量 $\widetilde{\boldsymbol{F}}$
8. FOR $\forall i=1,\cdots,n$ DO
9. 计算 $\widetilde{\boldsymbol{H}}(i,:)=\widetilde{\boldsymbol{F}}(i,:)/\sqrt{\mathrm{d}_{ii}}$
10. END FOR
11. 对 $\widetilde{\boldsymbol{H}}$ 进行标准的 k 均值聚类，得到最终的聚类结果

下面对上述算法的时间复杂度进行分析。第 1 步，求二部图矩阵需要计算每个顶点和每个锚点的欧氏距离，其时间复杂度为 $O(mnd)$，其中 d 为数据维度；第 2 步，对锚点集之间的图矩阵进行特征分解的时间复杂度为 $O(m^3)$；第 3 步，计算顶点集近似的度向量，要进行多次矩阵乘法，其最高的时间复杂度为 $O(nm)$；第 4 步，计算 $\widetilde{\boldsymbol{D}}^{-1/2}\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$ 需要先进行两次矩阵乘法，再按行进行除法，上述时间复杂度为 $O(nm)$；第 5 步，对维度为 $n\times m$ 的矩阵进行奇异值分解，相应的时间复杂度为 $O(nm^2)$；第 6 步，计算聚类指示矩阵，需要逐行进行除法，时间复杂度为 $O(nm)$；第 7 步，执行标准的 k 均值聚类算法，时间复杂度为 $O(nmkT)$，其中 T 为 k 均值聚类算法迭代次数。可以看出，在上述过程中，当样本 n 远远大于 m、d、T、k 时，整个算法的时间复杂度与 n 是线性相关的，因此该方法可以应用于大规模数据集。

4.2 二部图谱聚类算法的泛化分析

这一部分对算法 1 的泛化性进行分析。可知算法 1 近似的 Laplacian 矩阵有如下形式

$$\frac{1}{n}\widetilde{\boldsymbol{L}}_n=\frac{1}{n}(\boldsymbol{I}_n-\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}) \tag{15}$$

类似于第 3 节的描述，先给出上述 Laplacian 矩阵对应的经验算子 $\widetilde{L}_n$。可以看出在算法 1 中，标准 NCut 的相似度函数是由

$$\widetilde{W}(\boldsymbol{x},\boldsymbol{y})=\sum_{i,j=1}^{m}W(\boldsymbol{x},\boldsymbol{x}_i)\widetilde{K}_{ij}W(\boldsymbol{x}_j,\boldsymbol{y}) \tag{16}$$

所近似的，其中 $\widetilde{K}_{ij}$ 是位于矩阵 $\boldsymbol{K}^{-1}$ 的第 i 行第 j 列的元素。相应地，算法 1 的标准化的相似度矩阵为

$$\widetilde{G}_n(\boldsymbol{x},\boldsymbol{y})=\frac{\widetilde{T}(\boldsymbol{x},\boldsymbol{y})}{\sqrt{\widetilde{d}_n(\boldsymbol{x})\widetilde{d}_n(\boldsymbol{y})}} \tag{17}$$

其中 $\widetilde{d}_n(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}\widetilde{W}(\boldsymbol{x},\boldsymbol{x}_i)$。据此，可知算法 1 对应的经验 Laplacian 算子为

$$\widetilde{L}_nf(\boldsymbol{x})=f(\boldsymbol{x})-\int_X\widetilde{G}_n(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\rho_n(\boldsymbol{y}) \tag{18}$$

因此，$\frac{1}{n}\widetilde{\boldsymbol{L}}_n$ 的特征向量和 $\widetilde{\boldsymbol{L}}_n$ 的特征函数，有如下对应关系

$$\forall j\in[n],\widetilde{f}_{jt}=\widetilde{f}_t(\boldsymbol{x}_j)/\sqrt{n},$$
$$\forall\boldsymbol{x}\in X,\widetilde{f}_t(\boldsymbol{x})=\frac{\sum_{j=1}^{n}\widetilde{G}_n(\boldsymbol{x},\boldsymbol{x}_j)\widetilde{f}_t(\boldsymbol{x}_j)}{n\widetilde{\sigma}_t^2} \tag{19}$$

其中 $\widetilde{\sigma}_t$ 为 $\widetilde{\boldsymbol{D}}^{-1/2}\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1/2}$ 第 t 大的奇异值。

与此同时，相应的聚类指示函数有如下形式

$$\widetilde{H}=\{\widetilde{h}_t\}_{t=1}^k,\text{其中}\widetilde{h}_t(\boldsymbol{x})=\widetilde{f}_t(\boldsymbol{x})/\sqrt{\widetilde{d}_n(\boldsymbol{x})} \tag{20}$$

依照定义 4 所述,本节给出算法 1 得到的聚类指示函数的泛化风险 $R(\widetilde{H})-\widehat{R}(\widetilde{H})$ 和额外风险 $R(\widetilde{H})-R(H^*)$ 的上界,分别如定理 2 和定理 3 所示。

定理 2. 由算法 1、式 (19) 和式 (20) 得到的聚类指示函数 $\widetilde{H}$ 的泛化风险上界为 $\widetilde{O}(k/\sqrt{n})$。

注释. 上述泛化风险界说明,基于 Nyström 方法的二部图谱聚类算法具备较好的泛化性,并不会因为是对 NCut 的聚类指示函数的近似而影响其在整个空间上的表现。据此,本文给出了一种能够快速获取训练集外顶点外的低维特征的方法,该方法可以避免重复的奇异值分解,如算法 2 所示。该定理的证明参见附录 3。

算法 2. 快速计算训练集外顶点 k 维嵌入的算法

输入:通过算法 1 得到的前 k 个奇异值 $\{\widetilde{\sigma}_t\}_{t=1}^k$,对应的左奇异向量 $\{\boldsymbol{f}_t\}_{t=1}^k$,相似度函数 $\widetilde{W}(\cdot,\cdot)$,标准化的相似度函数 $\widetilde{G}_n(\cdot,\cdot)$,空间中任意顶点 $\boldsymbol{x}$

输出:顶点 $\boldsymbol{x}$ 的 k 维嵌入

1. FOR $\forall t=1,\cdots,k$ DO
2. 计算

$$\widetilde{f}_t(\boldsymbol{x})=\frac{\sum_{j=1}^n \widetilde{G}_n(\boldsymbol{x},\boldsymbol{x}_j)\widetilde{f}_t(\boldsymbol{x}_j)}{n\widetilde{\sigma}_t^2},$$

其中 $\widetilde{f}_{jt}$ 为 $\boldsymbol{f}_t$ 第 j 个分量

3. 计算近似的度函数

$$\widetilde{d}_n(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^n \widetilde{W}(\boldsymbol{x},\boldsymbol{x}_i)$$

4. 计算近似的聚类指示函数

$$\widetilde{h}_t(\boldsymbol{x})=\widetilde{f}_t(\boldsymbol{x})/\sqrt{\widetilde{d}_n(\boldsymbol{x})}$$

5. END FOR
6. 返回:顶点 $\boldsymbol{x}$ 的 k 维嵌入 $[\widetilde{h}_1(\boldsymbol{x}),\cdots,\widetilde{h}_k(\boldsymbol{x})]$

定理 3. 设 γ 是一个正数且 $\gamma\in[1,n]$,当从顶点集中均匀采样的锚点数为 $m=\Theta(\frac{n}{\gamma\varepsilon^2})$ 时,算法 1 所得的聚类指示矩阵的额外风险有上界如下

$$R(\widetilde{H})-R(H^*)\leqslant\widetilde{O}(\frac{k\gamma}{n(1-\varepsilon)}+\frac{k}{\sqrt{n}}) \quad (21)$$

注释. 从上述额外风险界可以看出,γ 越大,采样的锚点数就越少,于是额外风险界就会更大;随着 γ 变小,当 $\gamma=\Theta(\sqrt{n})$ 时,$m=\Theta(\sqrt{n}/\varepsilon^2)$,此时额外风险达到 $\widetilde{O}(k/\sqrt{n})$。因受标准 NCut 额外风险的影响(定理 1),当 $\gamma=o(\sqrt{n})$,上述风险界再也不会受 γ 的变小而变得更紧。此外,根据前面关于计算复杂度的分析,随着锚点变多,计算复杂度会更高。因此,可以看出,取 $\gamma=\Theta(\sqrt{n})$,即 $m=\Theta(\sqrt{n}/\varepsilon^2)$ 时,算法可以实现统计精度与计算效率的最佳平衡。该定理的证明参见附录 4。

5　实　验

本节主要通过实验证明所提推导的定理的正确性和所提出算法的有效性。在实验中,主要用到了四个大规模数据集:一是著名的 MNIST60K 手写数据集[37],其包含顶点数为 60000,实验将其分为 50000 个顶点组成的训练集和 10000 个顶点组成的测试集,该数据集维度为 784 维,共有 10 个类;二是 Winnipeg 数据集[38],其由光学和 PolSAR 遥感图像所组成,是从 Winnipeg 一个农业区附近收集的具有时间、光谱、纹理和极化特征等维度的农田数据集。Winnipeg 有 174 个特征维度,由 325834 个顶点组成,在实验过程中,这些顶点被分为了由 293254 个顶点组成的训练集和 32580 个顶点组成的测试集;三是 CIFAR10[39],其为一个图片数据集。它由 60000 个顶点组成,实验将其分为 50000 个顶点组成的训练集和 10000 个顶点组成的测试集;四是 EMNIST[40],是手写数据集 MNIST60K 的扩展,其有 240000 个顶点,实验将其分为由 216000 个顶点的训练集和 24000 个顶点组成的测试集。本实验将四个数据集的所有特征都进行了标准化处理,使其数值的绝对值都在区间 [0,1] 中。

5.1　二部图额外聚类风险上界的实验验证

第一个实验主要是用于验证定理 3 中所述的结论,亦即二部图谱聚类的额外聚类风险界与锚点数 m 之间的关系。由于无法获取数据集所属的整个顶点空间,因此,本文先在训练集上获取如算法 2 所述聚类指示函数 $\widetilde{H}$ (由式 (19) 和式 (20) 所得),再用这些聚类指示函数在测试集上的经验误差代替泛化误差,即

$$\widehat{R}_t(\widetilde{H})=\frac{1}{2n_t^2}\sum_{i,j=1}^{n_t}\sum_{t=1}^{k}W(\boldsymbol{z}_i,\boldsymbol{z}_j)(\widetilde{h}_t(\boldsymbol{z}_i)-\widetilde{h}_t(\boldsymbol{z}_j))^2 \quad (22)$$

其中 $\{\boldsymbol{z}_i\}_{i=1}^{n_t}$ 为测试集,n_t 为测试集顶点数。对于最优的泛化误差,本文直接在测试集上运行标准的 NCut 算法,获取其聚类指示函数 H^*,再计算出相应的最优泛化误差 $\widehat{R}_t(H^*)$。最后,利用 $\widehat{R}_t(\widetilde{H})-$

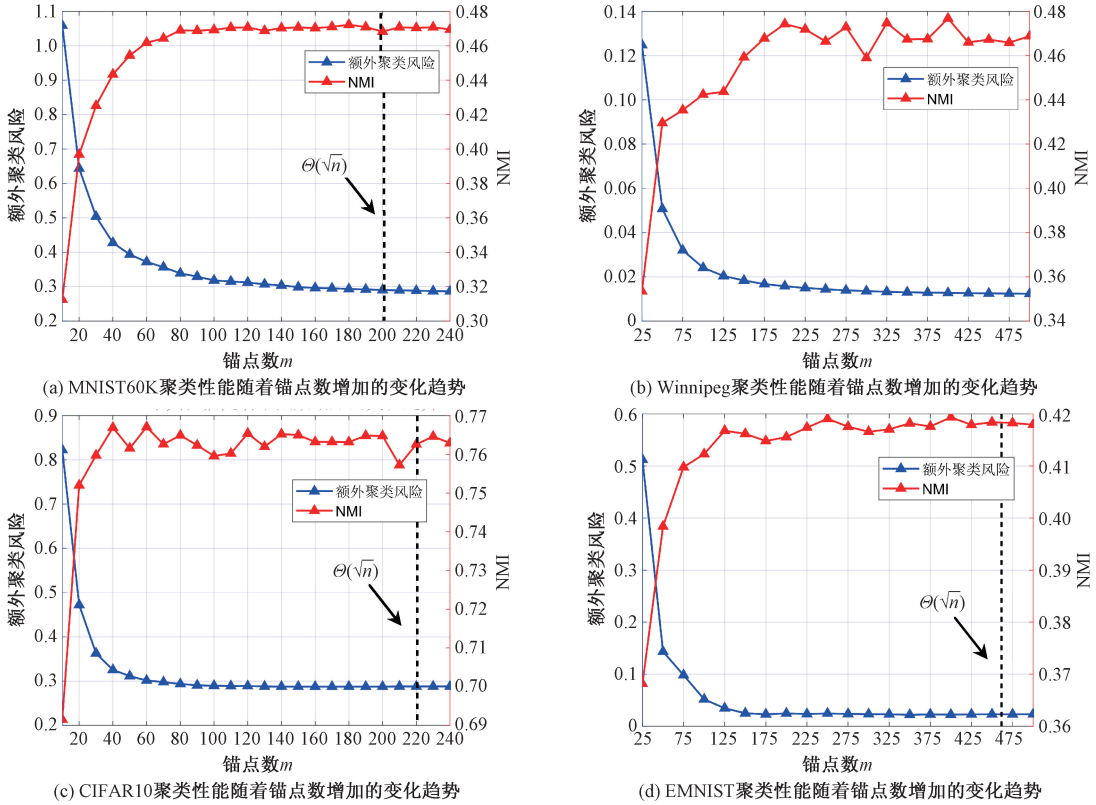
$\hat{R}_t(H^*)$来代替额外聚类风险。此外，本文还记录了相应的聚类结果随着锚点数增加而发生的变化。本文采用常用于评价聚类结果的标准化互信息(Normalized Mutual Information，NMI)。来验证算法1在训练集上的聚类表现，设样本真实标签为$Y=\{y_i\}_{i=1}^n$，算法的预测标签为$\hat{Y}=\{\hat{y}_i\}_{i=1}^n$，其具体的定义为

$$NMI(Y,\hat{Y})=\frac{2I(Y,\hat{Y})}{H(Y)+H(\hat{Y})} \tag{23}$$

其中$H(\cdot)$为交叉熵函数，$I(Y,\hat{Y})=H(Y)-H(\hat{Y}|Y)$为互信息函数。NMI取值范围在$[0,1]$中，且越接近于1，说明聚类效果越好。在实验中，参照文献[35]，实验中高斯核宽度参数σ^2取训练集顶点间距离平方的平均，即

$$\sigma^2=\frac{1}{n^2}\sum_{i,j=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2 \tag{24}$$

对于数据集MNIST60K和CIFAR10，本文使锚点数在{10:10:240}中取值；对于数据集Winnipeg和EMNIST，本文使锚点数在{25:25:500}中取值。为了减少随机性的影响，本文对相同的锚点数，采用不同随机种子，重复实验30次。在四个数据集上的实验结果如图1所示。图1蓝色曲线记录了算法1额外聚类风险随着锚点数增加的变化趋势，其值为30次实验的平均值。可以看出，随着m的增加，额外聚类风险下降非常快，呈现出$O\left(\frac{1}{m}\right)$的变化趋势，由于$m=\Theta\left(\frac{n}{\gamma\varepsilon^2}\right)$，这与定理3所述上界中的$\tilde{O}\left(\frac{k\varepsilon^2}{m(1-\varepsilon)}\right)$项基本吻合。为了更直观地反映聚类性能随锚点数增加的变化趋势，图1还将每个数据集等于$\Theta(\sqrt{n})$的锚点数位置用黑色虚线标出(其中数据集Winnipeg的$\sqrt{n}$超过500，故未标出)。从数据集MNIST60K的实验结果可以看出，当$m\approx\sqrt{n}$时，算法的额外聚类风险随着m的增加不再明显下降。而从其他三个数据集可以看出，在m明显小于$\sqrt{n}$时，算法的额外聚类风险便随着m的增加不再明显下降。上述观察充分说明了定理3的正确性。同时，从图1中各子图的红色曲线也可以看出聚类结果随着锚点数的增加逐渐变好，并在逐步趋于稳定。综上所述，二部图锚点数仅需取$\Theta(\sqrt{n})$即可达到令人满意的效果，这给相关二部图算法的锚点数选择提供了重要的参考依据。



(a) MNIST60K聚类性能随着锚点数增加的变化趋势

(b) Winnipeg聚类性能随着锚点数增加的变化趋势

(c) CIFAR10聚类性能随着锚点数增加的变化趋势

(d) EMNIST聚类性能随着锚点数增加的变化趋势

图1 在四个数据集上，算法1的额外聚类风险和NMI随着锚点数增加的变化趋势

5.2　算法 2 的实验验证

第二个实验主要用于验证算法 2 对于获取训练集外顶点的低维嵌入的高效性和有效性。本实验先在测试集上获取算法 2 所需的参数,将测试集视为训练集外的顶点,并通过算法 2 获取其低维嵌入,并记录算法所消耗的时间,以及利用这些低维嵌入获取的聚类结果。同时,本文还做了相关对比实验,即将测试集的顶点重新代入算法 1,用以获取相应的低维嵌入,同时记录相应的运行时间和聚类结果。

根据第一个验证额外风险上界的实验结果,在第二个实验中,锚点数取 $\lceil\sqrt{n}\rceil$,并重复实验 30 次取平均值。四个数据集的实验结果分别记录在表 2 至表 5 中。

表 2　算法 2 在 MNIST60K 测试集上的聚类效果和运行时间

	快速嵌入(算法 2)	标准嵌入(算法 1)	加速比率
NMI	0.4827±0.0091	0.4872±0.0081	
时间(s)	0.6445±0.1886	2.9187±1.6383	×4.52

表 3　算法 2 在 Winnipeg 测试集上的聚类效果和运行时间

	快速嵌入(算法 2)	标准嵌入(算法 1)	加速比率
NMI	0.4834±0.0183	0.4731±0.0151	
时间(s)	3.2314±0.1426	13.8387±0.5992	×4.28

表 4　算法 2 在 CIFAR10 测试集上的聚类效果和运行时间

	快速嵌入(算法 2)	标准嵌入(算法 1)	加速比率
NMI	0.7927±0.0148	0.8023±0.0115	
时间(s)	0.3800±0.0873	1.3050±0.2441	×3.43

表 5　算法 2 在 EMNIST 测试集上的聚类效果和运行时间

	快速嵌入(算法 2)	标准嵌入(算法 1)	加速比率
NMI	0.4331±0.0037	0.4336±0.0057	
时间(s)	3.2562±0.3159	13.8659±0.9599	×4.25

从表 2、表 3、表 4 和表 5 可以看出,算法 2 所提出的快速嵌入法只需要更少的时间即可获取聚类效果相似的训练集外顶点的低维嵌入,比标准嵌入方法要快数倍。这充分说明了算法 2 的有效性和高效性。

6　本文的局限与展望

本文给出了锚点为均匀采样的情况下,基于 Nyström 方法的二部图谱聚类泛化界。均匀采样是最常用的采样手段,然而该方法具有一定的"盲目性",在一定场景下会失效。例如,当类别不均衡时,有一些类顶点数量比较少,均匀采样会可能导致锚点集不含该类别的顶点。在这种情况下,就要用到基于重要性采样的手段[41-45]。这类方法旨在根据每个样本计算出其岭杠杆得分,具体如下

$$l(\boldsymbol{x}_i)=(\boldsymbol{W}(\boldsymbol{W}+\lambda n\boldsymbol{I}_n)^{-1})_{ii} \tag{25}$$

其中 λ 为超参数。上述岭杠杆得分的取值范围在 $[0,1]$ 之间,是顶点重要性的一种反映。接着,令采样概率为

$$p(\boldsymbol{x}_i)=\frac{l(\boldsymbol{x}_i)}{\sum_{i=1}^{n} l(\boldsymbol{x}_i)} \tag{26}$$

并依照上述概率采样 m 个锚点,可以采样具备更好统计性质的锚点集。然而,利用式(25)进行岭杠杆得分的计算需要得到完整的图矩阵并进行求逆的运算,这利用锚点进行加速的思想背道而驰。因此,文献[41-45]均采样了不同的方式,对岭杠杆得分进行了近似计算,以较低的复杂度得到了近似的岭杠杆得分。上述方法在核学习算法以及矩阵近似中取得了非常好的效果。然而,在二部图谱聚类中,上述锚点选择法能否取得更好的效果,即用更少的锚点数取得更好近似效果,尚属未被研究的问题。

此外,本文所研究的方法仅能处理单个视图,而现实中的数据往往具备多个视图,例如:建筑图纸中同一个建筑有正视图、俯视图和侧视图等三个角度的描述;同一个网页有网址、文本和视频等不同模态和视图的描述。为了处理多视图数据,研究者基于二部图谱聚类开发了大规模多视图谱聚类算法[46-51],取得了巨大的成功。在这类算法中,往往会引入视图系数,因此也会有不同的优化方法[49-50],这会对此类方法的泛化分析造成较大的困难。如何在考虑视图系数以及不同优化方法的情况下,对基于二部图的多视图谱聚类进行泛化分析,也是未来值得研究的一个方向。

7　总　结

本文先是建立了谱聚类泛化分析的框架。基于谱聚类一致性的性质,本文对标准 NCut 谱聚类算法进行了泛化分析,证明了其泛化风险上界和额外风险上界均为 $\widetilde{O}(k/\sqrt{n})$。接着,本文对基于 Nyström 方法的二部图谱聚类进行了泛化分析,并得到了其关于锚点数规模的泛化上界。基于上述理论结果,本文给出了一种最优锚点规模的选择方案,

即锚点规模为$\Theta(\sqrt{n})$时，二部图谱聚类算法达到统计精度和计算效率的最佳平衡。此外，本文还提出了一种能够快速计算训练集外顶点低维嵌入的算法。最后，本文还进行了综合实验，实验结果分别证明了上述理论的正确性以及算法的有效性。

致 谢 感谢国家自然科学基金面上项目(No.62276271)和国家自然科学基金杰出青年科学基金项目(No.62325604)的资助。

参 考 文 献

[1] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm//Advances in Neural Information Processing Systems. Vancouver, Canada, 2001, 14: 849-856

[2] Zeng S, Huang R, Kang Z, et al. Image segmentation using spectral clustering of gaussian mixture models. Neurocomputing, 2014, 144: 346-356

[3] Liu H Q, Jiao L C, Zhao F. Non-local spatial spectral clustering for image segmentation. Neurocomputing, 2010, 74(1-3): 461-471

[4] Mondal A, Giraldo J H, Bouwmans T, et al. Moving object detection for event-based vision using graph spectral clustering//Proceedings of the International Conference on Computer Vision Workshops. Montreal, Canada, 2021: 876-884

[5] Shin G, Albanie S, Xie W. Unsupervised salient object detection with spectral cluster voting//Proceedings of the Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 3971-3980

[6] Li Y, He K, Kloster K, et al. Local spectral clustering for overlapping community detection. ACM Transactions on Knowledge Discovery from Data, 2018, 12(2): 1-27

[7] Liu F, Choi D, Xie L, et al. Global spectral clustering in dynamic networks. Proceedings of the National Academy of Sciences, 2018, 115(5): 927-932

[8] Yu Z, Li L, You J, et al. Triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profiles. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(6): 1751-1765

[9] Yin L, Liu Y. Ensemble biclustering gene expression data based on the spectral clustering. Neural Computing and Applications, 2018, 30: 2403-2416

[10] Wang H, Lee T, Leung C C, et al. Acoustic segment modeling with spectral clustering methods. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(2): 264-277

[11] Bach F R, Jordan M I. Learning spectral clustering, with application to speech separation. The Journal of Machine Learning Research, 2006, 7: 1963-2001

[12] Liang W, Zhou S, Xiong J, et al. Multi-view spectral clustering with high-order optimal neighborhood laplacian matrix. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(7): 3418-3430

[13] Chen Man-Sheng, Cai Xiao-Sha, Lin Jia-Qi, et al. Tensor learning induced multi-view spectral clustering. Chinese Journal of Computers, 2024, 47(1):52-68 (in Chinese)
(陈曼笙,蔡晓莎,林家祺等.张量学习诱导的多视图谱聚类.计算机学报, 2024, 47(1):52-68)

[14] Tan Y, Liu Y, Wu H, et al. Euclidean distance is not your swiss army knife. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(12): 8179-8191

[15] Ren Y, Chen X, Xu J, et al. A novel federated multi-view clustering method for unaligned and incomplete data fusion. Information Fusion, 2024, 108: 102357

[16] Ling Y, Chen J, Ren Y, et al. Dual label-guided graph refinement for multi-view graph clustering//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(7): 8791-8798

[17] Chen X, Cai D. Large scale spectral clustering with landmark-based representation//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011, 25(1): 313-318

[18] Liu W, He J, Chang S F. Large graph construction for scalable semi-supervised learning//Proceedings of the International Conference on Machine Learning. Haifa, Israel, 2010: 679-686

[19] Liu J, Wang C, Danilevsky M, et al. Large-scale spectral clustering on graphs//Proceedings of the International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 1486-1492

[20] Fowlkes C, Belongie S, Chung F, et al. Spectral grouping using the Nyström method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214-225

[21] Hairer E, Wanner G. A theory for Nyström methods. Numerische Mathematik, 1975, 25: 383-400

[22] Williams C, Seeger M. Using the Nyström method to speed up kernel machines//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 2000, 13: 682-688

[23] Sun S, Zhao J, Zhu J. A review of Nyström methods for large-scale machine learning. Information Fusion, 2015, 26: 36-48

[24] Zhou Z, Amini A A. Analysis of spectral clustering algorithms for community detection: The general bipartite setting. Journal of Machine Learning Research, 2019, 20(47): 1-47

[25] Gu X, Deng J D. A multi-feature bipartite graph ensemble for image segmentation. Pattern Recognition Letters, 2020, 131: 98-104

[26] Yang X, Yu W, Wang R, et al. Fast spectral clustering learning with hierarchical bipartite graph for large-scale data. Pattern Recognition Letters, 2020, 130: 345-352

[27] Zhang Z, Xing F, Wang H, et al. Revisiting graph construction for fast image segmentation. Pattern Recognition, 2018, 78: 344-357

[28] Biau G, Devroye L, Lugosi G. On the performance of clustering in hilbert spaces. IEEE Transactions on Information Theory, 2008, 54(2): 781-790

[29] Liu Y. Refined learning bounds for kernel and approximate-means//Advances in Neural Information Processing Systems. Virtual, 2021, 34: 6142-6154

[30] Bucarelli M S, Larsen M, Schwiegelshohn C, et al. On generalization bounds for projective clustering//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2024, 36: 71723-71754

[31] Li S, Ouyang S, Liu Y. Understanding the generalization performance of spectral clustering algorithms//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(7): 8614-8621

[32] Von Luxburg U, Belkin M, Bousquet O. Consistency of spectral clustering. The Annals of Statistics, 2008, 36(2): 555-586

[33] Hoeffding W. Probability inequalities for sums of bounded random variables. The Collected Works of Wassily Hoeffding. New York: Springer, 1994

[34] Drineas P, Mahoney M W, Cristianini N. On the Nyström method for approximating a gram matrix for improved kernel-based learning. Journal of Machine Learning Research, 2005, 6(12): 2153-2175

[35] Wang S, Gittens A, Mahoney M W. Scalable kernel K-Means clustering with Nyström approximation: Relative-error bounds. Journal of Machine Learning Research, 2019, 20(12): 1-49

[36] Calandriello D, Rosasco L. Statistical and computational trade-offs in kernel K-Means//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018, 31: 9379-9389

[37] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[38] Crop Mapping Using Fused Optical-Radar Data Set. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/c5g89d

[39] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Master's Thesis, University of Toronto, Toronto, Canada, 2009

[40] Cohen G, Afshar S, Tapson J, et al. EMNIST: Extending MNIST to handwritten letters//Proceedings of the International Joint Conference on Neural Networks. Anchorage, USA, 2017: 2921-2926

[41] Huang L, Vishnoi N K. Coresets for clustering in euclidean spaces: Importance sampling is nearly optimal//Proceedings of the Annual ACM SIGACT Symposium on Theory of Computing. Chicago, USA, 2020: 1416-1429

[42] Chen Y, Yang Y. Fast statistical leverage score approximation in kernel ridge regression//Proceedings of the International Conference on Artificial Intelligence and Statistics. Virtual, 2021: 2935-2943

[43] Drineas P, Magdon-Ismail M, Mahoney M W, et al. Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research, 2012, 13(1): 3475-3506

[44] Cohen M B, Musco C, Musco C. Input sparsity time low-rank approximation via ridge leverage score sampling//Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms. Barcelona, Spain, 2017: 1758-1777

[45] Rudi A, Calandriello D, Carratino L, et al. On fast leverage score sampling and optimal learning//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018, 31: 5677-5687

[46] Li Y, Nie F, Huang H, et al. Large-scale multi-view spectral clustering via bipartite graph//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA 2015, 29(1): 2750-2756

[47] Wang Song-Gui, Wu Mi-Xia, Jia Zhong-Zhen. Matrix Inequality. Beijing: Science Press, 2006 (in Chinese)
(王松桂, 吴密霞, 贾忠贞. 矩阵不等式. 北京: 科学出版社, 2006)

[48] Li L, He H. Bipartite graph based multi-view clustering. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(7): 3111-3125

[49] Yang B, Zhang X, Nie F, et al. Fast multiview clustering with spectral embedding. IEEE Transactions on Image Processing, 2022, 31: 3884-3895

[50] Wang Y, Zhang W, Wu L, et al. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering//Proceedings of the International Joint Conference on Artificial Intelligence. New York, USA, 2016: 2153-2159

[51] Wang Y. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(1s): 1-25

[52] Rosasco L, Belkin M, De Vito E. On learning with integral operators. Journal of Machine Learning Research, 2010, 11(2): 905-934

[53] Cristianini N, Shawe-Taylor J. Kernel Methods for Pattern Analysis. Cambridge,UK: Cambridge University Press, 2004

附录 1. 预备知识

为了更好地阐述文中定理的证明过程，先对一些基础的数学知识进行介绍，主要有集中不等式、算子基础理论、核函数理论和谱聚类一致性等四个方面。

集中不等式。集中不等式描述了随机变量是否集中在某个固定值附近。著名的集中不等式包括 Markov 不等式、Chebyshev 不等式、Hoeffding 不等式[33]和 Bernstein 不等式等等。这些不等式在不同的领域都有重要应用，尤其在统计学习理论领域，起到了非常关键的作用。本文主要探讨的是经验聚类损失和期望聚类损失之间的差异，在证明过程中，主要应用的集中不等式是 Hoeffding 不等式，其具体描述如下。

定理 4（Hoeffding 不等式[33]）. 设 $\{z_i\}_{i=1}^n$ 为 n 个独立同分布的一维随机变量，且对于每个 z_i，都有 $z_i \in [a,b]$。设 $\mu = E[z_i]$，那么对于任意的 $\varepsilon > 0$，都有

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} z_i - \mu > \varepsilon\right) \leqslant \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \tag{27}$$

$$\Pr\left(\mu - \frac{1}{n}\sum_{i=1}^{n} z_i > \varepsilon\right) \leqslant \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \tag{28}$$

算子基础理论。在此，对泛函分析中的算子理论进行简要的介绍。算子是定义在函数空间上的运算，最常见的就是线性算子。当有限维空间的维度趋向于无穷大时，该空间可以被看作是函数空间。换言之，函数可以看作是维度无穷大的向量，而 n 维向量中的元素则可以看作是函数在 n 个点上的取值。比如，对于某个向量 $[f(x_1),\cdots,f(x_n)]^{\mathrm{T}}$，函数 f 可以看作是该向量在函数空间上的延拓，而该向量可以看作是函数 f 在 n 个点上的限定。

于是，相似度矩阵 $\frac{1}{n}\boldsymbol{W}$ 可以看作是 $\mathcal{R}^n \to \mathcal{R}^n$ 上的一个线性算子。设 $\frac{1}{n}\boldsymbol{W}$ 作用在向量 $[f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_n)]^{\mathrm{T}}$ 上，则有

$$\frac{1}{n}\boldsymbol{W}[f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_n)]^{\mathrm{T}} = \left[\frac{1}{n}\sum_{i=1}^{n} W_{1i}f(\boldsymbol{x}_i),\cdots,\frac{1}{n}\sum_{i=1}^{n} W_{ni}f(\boldsymbol{x}_i)\right]^{\mathrm{T}} \tag{29}$$

其中 W_{ji} 为矩阵 $\boldsymbol{W}$ 第 j 行第 i 列的元素，即 $W(\boldsymbol{x}_j,\boldsymbol{x}_i)$。当 $n \to \infty$ 时，训练集 $\{\boldsymbol{x}_i\}_{i=1}^n$ 延拓成了全空间 X，上式第 j 个元素可以写成

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} W_{ji}f(\boldsymbol{x}_i) = \int_X W(\boldsymbol{x}_j,\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x}) \tag{30}$$

可以看出，当 $n \to \infty$，$\frac{1}{n}\boldsymbol{W}$ 变成了连续函数空间 $C(X)$ 中的积分算子 L_w，其定义如下

$$L_w : C(X) \to C(X), L_w f(\boldsymbol{y}) = \int_X W(\boldsymbol{y},\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x}) \tag{31}$$

将式(31)中在 X 空间上的积分限定在训练集 $\{\boldsymbol{x}_i\}_{i=1}^n$ 上时，则可定义如下经验算子

$$\hat{L}_w : C(X) \to C(X), \hat{L}_w f(\boldsymbol{y}) = \int_X W(\boldsymbol{y},\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\rho_n(\boldsymbol{x}) \tag{32}$$

其中 ρ_n 表示与 $\{\boldsymbol{x}_i\}_{i=1}^n$ 有关的经验分布。

算子 L_w，$\hat{L}_w$ 和 $\frac{1}{n}\boldsymbol{W}$ 之间的特征函数和特征值有着广泛的联系，关于上述三个算子的谱性质可以参考文献[32]和文献[52]。

核函数理论。本文主要用高斯核函数作为顶点间的相似度函数，因此用到一些核函数的基本理论，具体如下。

定理 5（核函数的性质[53]）. 存在某个 Hilbert 空间 $\mathcal{H}$，使得定义在 X 上的核函数 $W(\cdot,\cdot): X \times X \to \mathcal{R}$ 可以写成如下 $\mathcal{H}$ 上内积的形式

$$W(\boldsymbol{x},\boldsymbol{y}) = \langle\varphi(\boldsymbol{x}),\varphi(\boldsymbol{y})\rangle_{\mathcal{H}} \tag{33}$$

其中 $\phi(\boldsymbol{x})$ 为 $\mathcal{H}$ 上的元素，因此 $\phi(\boldsymbol{x})$ 又被称为 $\boldsymbol{x}$ 的特征映射。

谱聚类一致性。下面对 NCut 的特征值和特征函数的一致性进行介绍，主要由如下两个定理给出。为了描述简洁，本文仅截取了文献[32]中定理 15 和定理 16 的关键部分。

定理 6（NCut 特征值一致性，文献[32]中定理 15） 设 $\mu(K)$ 为某个算子 K 的特征值组成的集合。令 λ 为期望 Laplacian 算子 L 的特征值，M 为 λ 的某个开邻域，使得 $M \cap \mu(K) = \lambda$。那么对于序列 $(\lambda_n)_{n\in N}$，且 $\lambda_n \in \mu(L_n) \cap M$，有 $\lambda_n \xrightarrow{a.s.} \lambda$。

定理 7（NCut 特征函数一致性，文献[32]中定理 16 及例 1） 在上述定理中，设 λ_n 对应的特征函数为 u_n，λ 对应的特征函数为 u，那么存在序列 $(a_n)_{n\in N}$，其中 $a_n \in \{-1,1\}$ 使得

$$\|a_n u_n - u\|_\infty \leqslant \frac{c}{\sqrt{n}} \tag{34}$$

其中 c 为某个大于 0 的常数。

附录 2. 定理 1 的证明

定理 1 的证明将分为以下两部分,第一部分证明了泛化风险的上界,第二部分则是关于额外风险的上界。

证明.(泛化风险上界)本文将多次利用 Hoeffding 不等式证明此定理,首先对 $\hat{h}_t$ 的有界性进行讨论。由于期望 Laplacian 算子的某个特征函数 f_t^* 在空间 χ 上满足单位约束,即 $\int (f_t^*(\boldsymbol{x}))^2 \mathrm{d}\rho(\boldsymbol{x}) = 1$。因此,对于任意的 $\boldsymbol{x} \in \chi$,$|f_t^*(\boldsymbol{x})|$ 必然有界。另外,随着 $n \to \infty$,$\hat{f}_t \to f_t^*$,可知 $\hat{f}_t$ 在空间 χ 上也是有界的.因此,$\hat{h}_t(\boldsymbol{x}) = \hat{f}_t(\boldsymbol{x}) / \sqrt{d_n(\boldsymbol{x})}$ 必然有界,设其上界为 b。

由于 $W(\cdot,\cdot)$ 具备对称性,可对 $\hat{R}(\hat{H})$ 进行如下分解

$$\hat{R}(\hat{H}) = \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}\sum_{t=1}^{k} W(\boldsymbol{x}_i,\boldsymbol{x}_j)\hat{h}_t^2(\boldsymbol{x}_i)}_{\mathcal{A}_1} - \underbrace{\frac{1}{n^2}\sum_{i,j=1}^{n}\sum_{t=1}^{k} W(\boldsymbol{x}_i,\boldsymbol{x}_j)\hat{h}_t(\boldsymbol{x}_i)\hat{h}_t(\boldsymbol{x}_j)}_{\mathcal{A}_2} \tag{35}$$

同理,可对 $R(\hat{H})$ 进行如下分解

$$R(\hat{H}) = \underbrace{\sum_{t=1}^{k}\iint W(\boldsymbol{x},\boldsymbol{y})\hat{h}_t^2(\boldsymbol{x})}_{\mathcal{B}_1} - \underbrace{\iint \sum_{t=1}^{k} W(\boldsymbol{x},\boldsymbol{y})\hat{h}_t(\boldsymbol{x})\hat{h}_t(\boldsymbol{y})}_{\mathcal{B}_2} \tag{36}$$

下面先给出 $\mathcal{B}_1 - \mathcal{A}_1$ 的上界。可知 $\mathcal{A}_1$ 可以进行如下表示 $\mathcal{A}_1 = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{k} d_n(\boldsymbol{x}_i)\hat{h}_t^2(\boldsymbol{x}_i)$。与此同时,可知 $\mathcal{B}_1 = \sum_{t=1}^{k}\int \mathrm{d}(\boldsymbol{x})\hat{h}_t^2(\boldsymbol{x})$。

对于任意的 $\boldsymbol{x},\boldsymbol{y} \in \chi$,由 $W(\boldsymbol{x},\boldsymbol{y})\hat{h}_t^2(\boldsymbol{x}) \leqslant b^2$,可由 Hoeffding 不等式知下式以 $1-\delta$ 的概率成立,

$$d(\boldsymbol{x})\hat{h}_t^2(\boldsymbol{x}) \leqslant d_n(\boldsymbol{x})\hat{h}_t^2(\boldsymbol{x}) + b^2\sqrt{\frac{\log(1/\delta)}{2n}} \tag{37}$$

由联合界,可知下式以 $1-\delta$ 的概率成立,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{k} d(\boldsymbol{x}_i)\hat{h}_t^2(\boldsymbol{x}_i) \leqslant \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{k} d_n(\boldsymbol{x}_i)\hat{h}_t^2(\boldsymbol{x}_i) + kb^2\sqrt{\frac{\log(kn/\delta)}{2n}} \tag{38}$$

此外,由 Hoeffding 不等式,可知下式以 $1-\delta$ 的概率成立,

$$\int \mathrm{d}(\boldsymbol{x})\hat{h}_t^2(\boldsymbol{x}) \leqslant \frac{1}{n}\sum_{i=1}^{n}\mathrm{d}(\boldsymbol{x}_i)\hat{h}_t^2(\boldsymbol{x}_i) + b^2\sqrt{\frac{\log(1/\delta)}{2n}} \tag{39}$$

因此,可知下式以 $1-\delta$ 的概率成立,

$$\sum_{t=1}^{k}\int \mathrm{d}(\boldsymbol{x})\hat{h}_t^2(\boldsymbol{x}) \leqslant \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{k} d(\boldsymbol{x}_i)\hat{h}_t^2(\boldsymbol{x}_i) + kb^2\sqrt{\frac{\log(kn/\delta)}{2n}} \tag{40}$$

结合式(38)和式(40),可知

$$\mathcal{B}_1 - \mathcal{A}_1 \leqslant 2kb^2\sqrt{\frac{\log(2kn/\delta)}{2n}} \tag{41}$$

成立的概率至少为 $1-\delta$。

下面给出 $\mathcal{B}_2 - \mathcal{A}_2$ 的上界。核函数可以表示为某个 Hilbert 空间中的内积。因此,对于任意两个顶点 $\boldsymbol{x},\boldsymbol{y} \in \chi$,有 $W(\boldsymbol{x},\boldsymbol{y}) = \langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{y}) \rangle$,其中 $\varphi(\cdot)$ 为 Hilbert 空间对应的特征映射。可知

$$\mathcal{A}_2 = \sum_{t=1}^{k}\left\langle \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x}_i)\hat{h}_t(\boldsymbol{x}_i), \frac{1}{n}\sum_{j=1}^{n}\varphi(\boldsymbol{x}_j)\hat{h}_t(\boldsymbol{x}_j)\right\rangle \tag{42}$$

以及

$$\mathcal{B}_2 = \sum_{t=1}^{k}\left\langle \int \varphi(\boldsymbol{x})\hat{h}_t(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x}), \int \varphi(\boldsymbol{y})\hat{h}_t(\boldsymbol{y})\mathrm{d}\rho(\boldsymbol{y})\right\rangle \tag{43}$$

对于 Hilbert 空间中任意向量 $\boldsymbol{u}$,满足 $\|\boldsymbol{u}\| \leqslant b$,易知 $\langle \varphi(\boldsymbol{x})\hat{h}_t(\boldsymbol{x}), \boldsymbol{u} \rangle \leqslant b^2$。由 Hoeffding 不等式可知,下式至少以 $1-\delta$ 的概率成立,

$$\left\langle \boldsymbol{u}, \int \varphi(\boldsymbol{x})\hat{h}_t(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x})\right\rangle \leqslant \left\langle \boldsymbol{u}, \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x}_i)\hat{h}_t(\boldsymbol{x}_i)\right\rangle + b^2\sqrt{\frac{\log(1/\delta)}{2n}} \tag{44}$$

由 $\boldsymbol{u}$ 的任意性,可得对于任意的 $t \in [k]$,

$$\left\langle \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x}_i)\hat{h}_t(\boldsymbol{x}_i), \int \varphi(\boldsymbol{x})\hat{h}_t(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x})\right\rangle \leqslant \left\langle \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x}_i)\hat{h}_t(\boldsymbol{x}_i), \frac{1}{n}\sum_{j=1}^{n}\varphi(\boldsymbol{x}_j)\hat{h}_t(\boldsymbol{x}_j)\right\rangle + b^2\sqrt{\frac{\log(1/\delta)}{2n}} \tag{45}$$

以至少 $1-\delta$ 的概率成立。

类似地,可知下式至少以 $1-\delta$ 的概率成立,

$$\left\langle \int \varphi(\boldsymbol{x})\hat{h}_t(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x}), \int \varphi(\boldsymbol{y})\hat{h}_t(\boldsymbol{y})\mathrm{d}\rho(\boldsymbol{y})\right\rangle \leqslant \left\langle \int \varphi(\boldsymbol{x})\hat{h}_t(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x}), \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x}_i)\hat{h}_t(\boldsymbol{x}_i)\right\rangle$$

$$+b^2\sqrt{\frac{\log(1/\delta)}{2n}}$$
$$=\langle\frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x}_i)\hat{h}_t(\boldsymbol{x}_i),\int\varphi(\boldsymbol{x})\hat{h}_t(\boldsymbol{x})\mathrm{d}\rho(\boldsymbol{x})\rangle$$
$$+b^2\sqrt{\frac{\log(1/\delta)}{2n}} \tag{46}$$

结合式(45)和式(46),由联合界,可知下式至少以 $1-\delta$ 的概率成立,

$$\mathcal{B}_2-\mathcal{A}_2\leqslant 2kb^2\sqrt{\frac{\log(2kn/\delta)}{2n}} \tag{47}$$

最后,结合式(41)和式(47),可知

$$R(\hat{H})-\hat{R}(\hat{H})\leqslant 4kb^2\sqrt{\frac{\log(4kn/\delta)}{2n}} \tag{48}$$

以至少 $1-\delta$ 的概率成立。

证明.(额外风险上界)设 $\hat{f}_t,f_t^*$ 均有上界为 b。根据文献[32]的定理 16 及例 1,可知存在常数 $c>0$,及关于 n 的数列 $(a_n)_n$,其中 $a_n\in\{1,-1\}$,有

$$\|a_n\hat{f}_t-f_t^*\|_\infty\leqslant\frac{c}{\sqrt{n}} \tag{49}$$

接着,对于任意的 $\boldsymbol{x}\in\mathcal{X}$,可知

$$\hat{h}_t^2(\boldsymbol{x})-(h_t^*(\boldsymbol{x}))^2$$
$$=\frac{\hat{f}_t^2(\boldsymbol{x})}{d_n(\boldsymbol{x})}-\frac{(f_t^*(\boldsymbol{x}))^2}{d(\boldsymbol{x})}$$
$$=\frac{\hat{f}_t^2(\boldsymbol{x})d(\boldsymbol{x})-(f_t^*(\boldsymbol{x}))^2d_n(\boldsymbol{x})}{d_n(\boldsymbol{x})d(\boldsymbol{x})}$$
$$\leqslant\frac{(\hat{f}_t^2(\boldsymbol{x})-(f_t^*(\boldsymbol{x}))^2)d(\boldsymbol{x})+(f_t^*(\boldsymbol{x}))^2(d(\boldsymbol{x})-d_n(\boldsymbol{x}))}{l^2} \tag{50}$$

由于

$$\hat{f}_t^2(\boldsymbol{x})-(f_t^*(\boldsymbol{x}))^2$$
$$\leqslant|a_n\hat{f}_t(\boldsymbol{x})-f_t^*(\boldsymbol{x})|\cdot|a_n\hat{f}_t(\boldsymbol{x})+f_t^*(\boldsymbol{x})|$$
$$\leqslant 2b\|a_n\hat{f}_t-f_t^*\|_\infty\leqslant\frac{2bc}{\sqrt{n}} \tag{51}$$

且由 Hoeffding 不等式,易知下式以至少 $1-\delta$ 的概率成立

$$d(\boldsymbol{x})-d_n(\boldsymbol{x})\leqslant\sqrt{\frac{\log(1/\delta)}{2n}} \tag{52}$$

将式(51)和式(52)代入式(50),可知下式以至少 $1-\delta$ 的概率成立

$$\hat{h}_t^2(\boldsymbol{x})-(h_t^*(\boldsymbol{x}))^2$$
$$\leqslant\frac{1}{l^2}\left(\frac{2bc}{\sqrt{n}}+b^2\sqrt{\frac{\log(1/\delta)}{2n}}\right) \tag{53}$$

此外,对于任意的 $\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}$,有

$$\hat{h}_t(\boldsymbol{x})\hat{h}_t(\boldsymbol{y})-h_t^*(\boldsymbol{x})h_t^*(\boldsymbol{y})$$
$$=\frac{\hat{f}_t(\boldsymbol{x})\hat{f}_t(\boldsymbol{y})}{\sqrt{d_n(\boldsymbol{x})d_n(\boldsymbol{y})}}-\frac{f_t^*(\boldsymbol{x})f_t^*(\boldsymbol{y})}{\sqrt{d(\boldsymbol{x})d(\boldsymbol{y})}}$$
$$=\frac{\hat{f}_t(\boldsymbol{x})\hat{f}_t(\boldsymbol{y})\sqrt{d(\boldsymbol{x})d(\boldsymbol{y})}-f_t^*(\boldsymbol{x})f_t^*(\boldsymbol{y})\sqrt{d_n(\boldsymbol{x})d_n(\boldsymbol{y})}}{\sqrt{d(\boldsymbol{x})d(\boldsymbol{y})d_n(\boldsymbol{x})d_n(\boldsymbol{y})}}$$
$$\leqslant\frac{(\hat{f}_t(\boldsymbol{x})\hat{f}_t(\boldsymbol{y})-f_t^*(\boldsymbol{x})f_t^*(\boldsymbol{y}))\sqrt{d(\boldsymbol{x})d(\boldsymbol{y})}}{l^2}$$
$$+\frac{f_t^*(\boldsymbol{x})f_t^*(\boldsymbol{y})(\sqrt{d(\boldsymbol{x})d(\boldsymbol{y})}-\sqrt{d_n(\boldsymbol{x})d_n(\boldsymbol{y})})}{l^2} \tag{54}$$

由于

$$\hat{f}_t(\boldsymbol{x})\hat{f}_t(\boldsymbol{y})-f_t^*(\boldsymbol{x})f_t^*(\boldsymbol{y})$$
$$=(a_n\hat{f}_t(\boldsymbol{x})-f^*(\boldsymbol{x}))(a_n\hat{f}_t(\boldsymbol{y}))$$
$$+f_t^*(\boldsymbol{x})(a_n\hat{f}_t(\boldsymbol{y})-f_t^*(\boldsymbol{y}))$$
$$\leqslant 2b\|\hat{f}_t-f_t^*\|_\infty\leqslant\frac{2bc}{\sqrt{n}} \tag{55}$$

此外,还可得

$$\sqrt{d(\boldsymbol{x})d(\boldsymbol{y})}-\sqrt{d_n(\boldsymbol{x})d_n(\boldsymbol{y})}$$
$$\leqslant\sqrt{|d(\boldsymbol{x})d(\boldsymbol{y})-d_n(\boldsymbol{x})d_n(\boldsymbol{y})|}$$
$$\leqslant\frac{|d(\boldsymbol{x})d(\boldsymbol{y})-d_n(\boldsymbol{x})d_n(\boldsymbol{y})|}{l}$$
$$=\frac{|(d(\boldsymbol{x})-d_n(\boldsymbol{x}))d(\boldsymbol{y})+d_n(\boldsymbol{x})(d(\boldsymbol{y})-d_n(\boldsymbol{y}))|}{l}$$
$$\leqslant\frac{2}{l}\sqrt{\frac{\log(2/\delta)}{2n}} \tag{56}$$

以至少 $1-\delta$ 的概率成立。

将式(55)和式(56)代入式(54),可知下式以至少 $1-\delta$ 的概率成立

$$\hat{h}_t(\boldsymbol{x})\hat{h}_t(\boldsymbol{y})-h_t^*(\boldsymbol{x})h_t^*(\boldsymbol{y})$$
$$\leqslant\frac{2bc}{l^2\sqrt{n}}+\frac{2b^2}{l^3}\sqrt{\frac{\log(2/\delta)}{2n}} \tag{57}$$

下面给出 $R(\hat{H})-R(H^*)$ 的上界,结合式(53)和式(57),以及联合界,可知

$$R(\hat{H})-R(H^*)$$
$$=\frac{1}{2}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\sum_{t=1}^{k}W(\boldsymbol{x},\boldsymbol{y})(\hat{h}_t(\boldsymbol{x})-\hat{h}_t(\boldsymbol{y}))^2\right]$$
$$-\frac{1}{2}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\sum_{t=1}^{k}W(\boldsymbol{x},\boldsymbol{y})(h_t^*(\boldsymbol{x})-h_t^*(\boldsymbol{y}))^2\right]$$
$$\leqslant\sup_{\boldsymbol{x}}\sum_{t=1}^{k}|\hat{h}_t^2(\boldsymbol{x})-(h_t^*(\boldsymbol{x}))^2|$$
$$+\sup_{\boldsymbol{x},\boldsymbol{y}}\sum_{t=1}^{k}|\hat{h}_t(\boldsymbol{x})\hat{h}_t(\boldsymbol{y})-h_t^*(\boldsymbol{x})h_t^*(\boldsymbol{y})|$$

$$\leqslant \frac{2k}{l^3}\left(\frac{4bc}{\sqrt{n}}+b^2\sqrt{\frac{\log(4k/\delta)}{2n}}\right) \tag{58}$$

以至少 $1-\delta$ 的概率成立。

附录 3. 定理 2 的证明

为了证明方便,引入如下符号。由于任意样本 $\boldsymbol{x}$ 在再生核 Hilbert 空间中可表示为 $\varphi(\boldsymbol{x})$,设 $\boldsymbol{\Phi}_n=[\varphi(\boldsymbol{x}_1),\cdots,\varphi(\boldsymbol{x}_n)]$,那么 $\boldsymbol{W}=\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{\Phi}_n$。设在 Hilbert 空间中,采样锚点的表示为 $\boldsymbol{\Phi}_m=[\varphi(\boldsymbol{a}_1),\cdots,\varphi(\boldsymbol{a}_m)]$。那么

$$\boldsymbol{P}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}=\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{\Phi}_m(\boldsymbol{\Phi}_m^{\mathrm{T}}\boldsymbol{\Phi}_m)^{-1}\boldsymbol{\Phi}_m^{\mathrm{T}}\boldsymbol{\Phi}_n \tag{59}$$

同时,令 $\boldsymbol{\Pi}_{\mathrm{m}}=\boldsymbol{\Phi}_m(\boldsymbol{\Phi}_m^{\mathrm{T}}\boldsymbol{\Phi}_m)^{-1}\boldsymbol{\Phi}_m^{\mathrm{T}}$ 为 m 个锚点在 Hilbert 空间中张成空间所对应的正交投影矩阵。同理,令 n 个顶点的正交投影为 $\boldsymbol{\Pi}_{\mathrm{n}}=\boldsymbol{\Phi}_n(\boldsymbol{\Phi}_n^{\mathrm{T}}\boldsymbol{\Phi}_n)^{-1}\boldsymbol{\Phi}_n^{\mathrm{T}}$。为了衡量 $\boldsymbol{\Pi}_{\mathrm{n}}$ 和 $\boldsymbol{\Pi}_{\mathrm{m}}$ 之间的差异,先引入下述引理。

引理 1[36]. 设 γ 为任意正实数,当随机均匀采样 m 个锚点,满足 $m=\widetilde{\Theta}\left(\frac{n}{\gamma\varepsilon^2}\right)$ 时,下列不等式以至少 $1-\delta$ 的概率成立

$$\boldsymbol{\Pi}_n-\boldsymbol{\Pi}_m\preccurlyeq\frac{\gamma}{1-\varepsilon}(\Phi_n\Phi_n^{\mathrm{T}}+\gamma\boldsymbol{\Pi}_n)^{-1} \tag{60}$$

在上述引理中,$\preccurlyeq$ 为偏序符号,其具体含义如下:对于矩阵 $\boldsymbol{A},\boldsymbol{B}\in\mathcal{R}^{n\times n}$,若对于任意向量 $\boldsymbol{x}\in\mathcal{R}^n$,都有 $\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{x}\leqslant 0$,则将 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 之间的关系表示为 $\boldsymbol{A}\preccurlyeq\boldsymbol{B}$。接下来完成定理 2 的证明。

证明. 类似于定理 1,定理 2 也可由 Hoeffding 不等式证明。因此,只要证明对于任意的 n,函数 $\widetilde{h}_t$ 有界,定理即得证。根据前面关于正交投影的描述,可知

$$\begin{aligned}&|\widetilde{W}(\boldsymbol{x},\boldsymbol{y})-W(\boldsymbol{x},\boldsymbol{y})|\\&=|\varphi^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{\Pi}_m\varphi(\boldsymbol{y})-\varphi^{\mathrm{T}}(\boldsymbol{x})\varphi(\boldsymbol{y})|\\&=|\varphi^{\mathrm{T}}(\boldsymbol{x})(\boldsymbol{\Pi}_m-\boldsymbol{\Pi}_n)\varphi(\boldsymbol{y})|\\&\leqslant\|\varphi(\boldsymbol{x})\|\cdot\|\boldsymbol{\Pi}_m-\boldsymbol{\Pi}_n\|\cdot\|\varphi(\boldsymbol{y})\|\end{aligned} \tag{61}$$

此外,对于再生核 Hilbert 空间中的任意 $\varphi(\boldsymbol{x})$,有

$$\varphi^{\mathrm{T}}(\boldsymbol{x})(\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}})\boldsymbol{\varphi}(\boldsymbol{x})=\sum_{i=1}^{n}(\varphi^{\mathrm{T}}(\boldsymbol{x})\varphi(\boldsymbol{x}_i))^2\geqslant nl^2 \tag{62}$$

由 $\varphi(\boldsymbol{x})$ 的任意性,可得

$$\|(\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}}+\gamma\boldsymbol{\Pi}_n)^{-1}\|\leqslant\|(\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}})^{-1}\|\leqslant\frac{1}{nl^2} \tag{63}$$

接着,由于 $\boldsymbol{\Pi}_m\preccurlyeq\boldsymbol{\Pi}_n$,且由引理 1,结合式 (61),可知

$$|\widetilde{W}(\boldsymbol{x},\boldsymbol{y})-W(\boldsymbol{x},\boldsymbol{y})|\leqslant\frac{\gamma}{nl^2(1-\varepsilon)} \tag{64}$$

由于 $\gamma\ll n$,当 n 足够大时,可知 $\widetilde{W}(\boldsymbol{x},\boldsymbol{y})$ 必然有界. 接下来证明 $1/\widetilde{d}_n(\boldsymbol{x})$ 是有界的。据定义,可知

$$\begin{aligned}&|\widetilde{d}_n(\boldsymbol{x})-d_n(\boldsymbol{x})|\\&=\frac{1}{n}\left|\sum_{i=1}^{n}\widetilde{W}(\boldsymbol{x},\boldsymbol{x}_i)-\sum_{i=1}^{n}W(\boldsymbol{x},\boldsymbol{x}_i)\right|\\&\leqslant\frac{1}{n}\sum_{i=1}^{n}|\widetilde{W}(\boldsymbol{x},\boldsymbol{x}_i)-W(\boldsymbol{x},\boldsymbol{x}_i)|\\&\leqslant\frac{\gamma}{nl^2(1-\varepsilon)}\end{aligned} \tag{65}$$

据此可知,$\widetilde{d}_n(\boldsymbol{x})\geqslant l-\frac{\gamma}{nl^2(1-\varepsilon)}$. 当 $\gamma\ll n$ 时,$1/\widetilde{d}_n(\boldsymbol{x})$ 必然是有界的。

此时,结合 $\widetilde{W}(\boldsymbol{x},\boldsymbol{y})$ 和 $1/\widetilde{d}_n(\boldsymbol{x})$ 的有界性

$$\widetilde{G}_n(\boldsymbol{x},\boldsymbol{y})=\frac{\widetilde{W}(\boldsymbol{x},\boldsymbol{y})}{\sqrt{\widetilde{d}_n(\boldsymbol{x})\widetilde{d}_n(\boldsymbol{y})}} \tag{66}$$

必然有界,设其上界为 b。

那么由 Cauchy-Swartz 不等式可知

$$\begin{aligned}|\widetilde{f}_t(\boldsymbol{x})|&=\left|\frac{\sum_{j=1}^{n}\widetilde{G}_n(\boldsymbol{x},\boldsymbol{x}_j)\widetilde{f}_{it}}{\sqrt{n}(1-\widetilde{\sigma}_t^2)}\right|\\&\leqslant\frac{\sqrt{\sum_{j=1}^{n}\widetilde{G}_n^2(\boldsymbol{x},\boldsymbol{x}_j)}\sqrt{\sum_{i=1}^{n}\widetilde{f}_{it}^2}}{\sqrt{n}(1-\widetilde{\sigma}_t^2)}\\&\leqslant\frac{b}{1-\widetilde{\sigma}_t^2}\end{aligned} \tag{67}$$

显然 $\widetilde{\sigma}_t^2$ 严格小于 1,所以 $\widetilde{f}_t(\boldsymbol{x})$ 必然有界。再结合 $1/\widetilde{d}_n(\boldsymbol{x})$ 有界,可知 $\widetilde{h}_t(\boldsymbol{x})=\widetilde{f}_t(\boldsymbol{x})/\sqrt{\widetilde{d}_n(\boldsymbol{x})}$ 必然有界。

附录 4. 定理 3 的证明

在证明定理 3 之前,先给出如下引理。

引理 2(文献[47]的定理 6.2.2). 设矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是两个 $n\times n$ 的 Hermite 矩阵,且 $0_n\preccurlyeq\boldsymbol{A},\boldsymbol{B}$,那么

$$0\leqslant tr(\boldsymbol{A}\boldsymbol{B})\leqslant\|\boldsymbol{A}\|tr(\boldsymbol{A})\leqslant tr(\boldsymbol{A})tr(\boldsymbol{B})。$$

定理 3 的证明. 可对 $R(\widetilde{H})-R(H^*)$ 进行如下分解

$$\begin{aligned}&R(\widetilde{H})-R(H^*)\\&=\underbrace{R(\widetilde{H})-\widehat{R}(\widetilde{H})}_{\mathcal{A}}+\underbrace{\widehat{R}(\widetilde{H})-\widehat{R}(\widehat{H})}_{\mathcal{B}}\\&\quad+\underbrace{\widehat{R}(\widehat{H})-R(\widehat{H})}_{\mathcal{C}}+\underbrace{R(\widehat{H})-R(H^*)}_{\mathcal{D}}\end{aligned} \tag{68}$$

由正文中的定理 2,可知 $\mathcal{A}$ 的上界为 $\widetilde{O}(k/\sqrt{n})$.

$\mathcal{C}$ 为 $\widehat{H}$ 泛化风险的相反数，根据 Hoeffding 不等式的性质，易知 $\mathcal{C}$ 与 $\widehat{H}$ 泛化风险有相同上界。由定理 1 可知，$\mathcal{C}$ 和 $\mathcal{D}$ 的上界均为 $\widetilde{O}(k/\sqrt{n})$。接下来给出 $\mathcal{B}$ 的上界。

将 $\mathcal{B}$ 中的项写成矩阵形式，易知

$$\mathcal{B}=\frac{1}{n}tr(\widetilde{\boldsymbol{F}}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{W})\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{F}})-\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{F}) \tag{69}$$

并作如下分解

$$\begin{aligned}&\frac{1}{n}tr(\widetilde{\boldsymbol{F}}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{W})\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{F}})-\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{F})\\&=\left.\begin{aligned}&\frac{1}{n}tr(\widetilde{\boldsymbol{F}}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}(\boldsymbol{D}-\boldsymbol{W})\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{F}})\\&-\frac{1}{n}tr(\widetilde{\boldsymbol{F}}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}(\widetilde{\boldsymbol{D}}-\widetilde{\boldsymbol{W}})\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{F}})\end{aligned}\right\}\varepsilon\\&\left.\begin{aligned}&+\frac{1}{n}tr(\widetilde{\boldsymbol{F}}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}(\widetilde{\boldsymbol{D}}-\widetilde{\boldsymbol{W}})\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{F}})\\&-\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}(\widetilde{\boldsymbol{D}}-\widetilde{\boldsymbol{W}})\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{F})\}\boldsymbol{F}\end{aligned}\right\}\mathcal{F}\\&\left.\begin{aligned}&+\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}(\widetilde{\boldsymbol{D}}-\widetilde{\boldsymbol{W}})\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{F})\\&-\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{D}-W)\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{F}).\end{aligned}\right\}\mathcal{G}\end{aligned}$$

首先给出 ε 的上界，根据矩阵迹的一个基本不等式（文献[47]定理 6.2.2），可知

$$\varepsilon\leqslant\frac{1}{n}tr(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}})\cdot\|\boldsymbol{D}-\boldsymbol{W}-\widetilde{\boldsymbol{D}}+\widetilde{\boldsymbol{W}}\| \tag{70}$$

还是根据文献[47]中的定理 6.2.2，可知

$$tr(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}})\leqslant\|\widetilde{\boldsymbol{D}}^{-1}\|\cdot tr(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}) \tag{71}$$

由于 $\widetilde{\boldsymbol{D}}^{-1}$ 中的元素均是有界的，且 $tr(\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}})=k$，所以

$$tr(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}})\leqslant O(k) \tag{72}$$

此外，由引理 1，可知

$$\begin{aligned}&\frac{1}{n}\|\boldsymbol{W}-\widetilde{\boldsymbol{W}}\|\\&=\frac{1}{n}\|\boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\Pi}_n-\boldsymbol{\Pi}_m)\boldsymbol{\Phi}_n\|\\&\leqslant\frac{\gamma}{n(1-\varepsilon)}\|\boldsymbol{\Phi}_n^{\mathrm{T}}(\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{\mathrm{T}}+\gamma\boldsymbol{\Pi}_n)^{-1}\boldsymbol{\Phi}_n\|\\&\leqslant\frac{\gamma}{n(1-\varepsilon)}\end{aligned} \tag{73}$$

接着，可知

$$\begin{aligned}&\frac{1}{n}\|\boldsymbol{D}-\widetilde{\boldsymbol{D}}\|=\frac{1}{n}\|\boldsymbol{W}\boldsymbol{1}_n-\widetilde{\boldsymbol{W}}\boldsymbol{1}_n\|_\infty\\&=\sup_{x\in X}|d_n(\boldsymbol{x})-\widetilde{d}_n(\boldsymbol{x})|\\&\leqslant\frac{\gamma}{nl^2(1-\varepsilon)}\end{aligned} \tag{74}$$

将式(72)、式(73)和式(74)中的上界代入式(70)中，可知

$$\varepsilon\leqslant O(\frac{k\gamma}{n(1-\varepsilon)}) \tag{75}$$

下面给出 $\mathcal{F}$ 的上界，根据 $\widetilde{F}$ 的性质，可知

$$\mathcal{F}\leqslant 0 \tag{76}$$

对于 $\mathcal{G}$，可知

$$\mathcal{G}=\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{F})-\frac{1}{n}tr(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{F}) \tag{77}$$

进一步可化为

$$\frac{1}{n}\sum_{i,j=1}^{n}\sum_{t=1}^{k}f_{it}f_{jt}(\frac{\widetilde{W}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)}}-\frac{W(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{d_n(\boldsymbol{x}_i)d_n(\boldsymbol{x}_j)}}) \tag{78}$$

对于任意的两个顶点 $\boldsymbol{x}_i,\boldsymbol{x}_j$，可知存在某个正的常数 c 使得

$$\begin{aligned}&\left|\frac{1}{\sqrt{\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)}}-\frac{1}{\sqrt{d_n(\boldsymbol{x}_i)d_n(\boldsymbol{x}_j)}}\right|\\&\leqslant\frac{c|\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)-d_n(\boldsymbol{x}_i)d_n(\boldsymbol{x}_j)|}{\sqrt{\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)}+\sqrt{d_n(\boldsymbol{x}_i)d_n(\boldsymbol{x}_j)}}\\&\leqslant\frac{c^2}{2}|\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)-d_n(\boldsymbol{x}_i)d_n(\boldsymbol{x}_j)|\\&\leqslant c^2\sup_{\boldsymbol{x}}|\widetilde{d}_n(\boldsymbol{x})-d_n(\boldsymbol{x})|\\&\leqslant O\Big(\frac{\gamma}{n(1-\varepsilon)}\Big)\end{aligned} \tag{79}$$

此外，由定理 2 中的证明，可知

$$|\widetilde{W}(\boldsymbol{x}_i,\boldsymbol{x}_j)-W(\boldsymbol{x}_i,\boldsymbol{x}_j)|\leqslant\frac{\gamma}{nl^2(1-\varepsilon)} \tag{80}$$

结合式(79)和式(80)，可知

$$\begin{aligned}&\left|\frac{\widetilde{W}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)}}-\frac{W(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{d_n(\boldsymbol{x}_i)d_n(\boldsymbol{x}_j)}}\right|\\&\leqslant\left|\frac{\widetilde{W}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)}}-\frac{W(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)}}\right|\\&+\left|\frac{W(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{\widetilde{d}_n(\boldsymbol{x}_i)\widetilde{d}_n(\boldsymbol{x}_j)}}-\frac{W(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{d_n(\boldsymbol{x}_i)d_n(\boldsymbol{x}_j)}}\right|\\&\leqslant O(\frac{\gamma}{n(1-\varepsilon)})\end{aligned} \tag{81}$$

将上述结果代入式(78)中,可知

$$\begin{aligned}\mathcal{G}&\leqslant O\left(\frac{\gamma}{n(1-\varepsilon)}\right)\cdot\frac{1}{n}\sum_{i,j=1}^{n}\sum_{t=1}^{k}|f_{it}f_{jt}|\\&\leqslant O\left(\frac{\gamma}{n(1-\varepsilon)}\right)\cdot\frac{1}{n}\sum_{i,j=1}^{n}\sum_{t=1}^{k}\frac{f_{it}^2+f_{jt}^2}{2}\\&=O\left(\frac{k\gamma}{n(1-\varepsilon)}\right)\end{aligned}\tag{82}$$

结合式(75)、式(76)和式(82)的上界,将它们代入式(69)中,可知

$$\mathcal{B}\leqslant O\left(\frac{k\gamma}{n(1-\varepsilon)}\right)\tag{83}$$

综上所述,可得最终的结论

$$R(\widetilde{H})-R(H^*)\leqslant\widetilde{O}\left(\frac{k\gamma}{n(1-\varepsilon)}+\frac{k}{\sqrt{n}}\right)\tag{84}$$

LIANG Wei-Xuan, Ph. D., postdoctoral. His research interests include multi-view clustering, kernel learning, and learning theory.

LIU Xin-Wang, Ph. D., professor. His research interests include multi-view clustering and kernel learning.

LAN Long, Ph. D., associate researcher. His research interests include object detection, object ReID and clustering algorithm.

ZHU En, Ph. D., professor. His research interests include clustering algorithm, anomaly detection and pattern recognition.

Background

The topic of this paper is the generalization analysis of spectral clustering on bipartite graph. Due to its ability to effectively handle non-linearly separable and large-scale data, spectral clustering on bipartite graph has been successfully applied in numerous learning tasks with excellent results, such as image segmentation, object detection, community detection, gene expression analysis, speech segmentation, and multi-view clustering.

In the existing literature, spectral clustering on bipartite graph still faces the following three issues: 1. A lack of necessary theoretical guarantees, with no analysis of the algorithm's generalization ability. 2. Difficulty in efficiently obtaining positional embeddings for out-of-sample vertices. 3. Challenges in determining the number of anchor points to achieve an optimal balance between statistical accuracy and computational cost.

To address these issues, this paper first establishes a framework for analyzing the generalization of spectral clustering and, based on the consistency of spectral clustering, derives the upper bounds of both the generalization risk and the excess risk of the standard NCut algorithm. Then, this paper also analyzes the generalization ability of an approximation algorithm for the standard NCut, i. e., the Nyström-based spectral clustering algorithm on bipartite graph. Based on the derived generalization theory for spectral clustering on bipartite graph, this paper proposes an algorithm that can obtain low-dimensional embeddings for out-of-sample points. Furthermore, a theoretical strategy for selecting the number of anchor points is proposed, revealing that when the number of anchor points is , the algorithm achieves the optimal trade-off between statistical accuracy and computational efficiency.

This research is supported by the National Natural Science Foundation of China (No. 62325604, 62276271). This paper is the first work to analyze the generalization ability of spectral clustering on bipartite graph. It reveals the impact of the number of anchor points on the performance of bipartite graph spectral clustering, which is significant for understanding the properties of spectral clustering algorithms and developing new ones.