arXiv:2303.01415v1 [math.AT] 2 Mar 2023

局部数据结构

J.F. Jardine*

jardine@uwo.ca

2023年3月3日

摘要

局部数据结构是数据集中的邻域系统。邻域的定义可以有多种来源，例如全局几何结构（星图）、组合结构（加权图）、预期的计算结果（自然语言处理）或抽样。本文将在邻域理论的背景下讨论这些例子。

这一理论是理解大型数据集聚类问题的一步。这些簇在实践中只能近似得到，但可以根据邻域，通过源自 Healy-McInnes UMAP 构造的拼接论证构造出近似结果。通过将数据集结构的理论基础从度量空间改为扩展伪度量空间，可以运用这些拼接论证。

引言

本文初步探讨大型数据集的局部结构。

潜在的研究对象包括子集 $\mathcal{U} \subset \mathbb{R}^N$，其中数据集 $\mathcal{U}$（或“全集”）本质上是无限的，也就是说，$\mathcal{U}$ 对于现有计算设备而言过于庞大，无法进行分析。

或者，数据集 $\mathcal{U}$上可能不存在度量空间结构。此类对象 $\mathcal{U}$ 可能作为大型加权图 Γ的顶点出现，这些图可以描述在某个时间区间内发生的数据传输。其他例子则出现在自然语言处理的“词袋”模型中；该模型具有一种非图论的组合结构。

最后，$\mathcal{U}$可能并不存在明显的几何或组合结构，其在某点附近的结构可能需要通过迭代采样来近似（或学习）。

通常，人们希望将数据集 $\mathcal{U}$拆分为较小的可计算部分N，这些部分能够覆盖 $\mathcal{U}$，也就是说，每个 $x \in \mathcal{U}$ 都位于某个邻域 N中，

*由加拿大自然科学与工程研究理事会资助。

并希望或期待能够将对这些邻域 N 的分析汇总起来，从而对通用数据集 $\mathcal{U}$进行完整分析，或至少得到有用的部分分析。这基本上就是 Mapper 算法 [6] 所采用的方法（见下文注12），对于在相对较小的距离尺度上进行聚类而言，这种方法完全合理。

邻域 N 中的元素应该在某种意义上彼此接近，但我们必须解决如何在一片数据 $\mathcal{U}$海洋中找到这样的邻域这一问题。如果事先不知道 $\mathcal{U}$的结构或生成方式，那么“接近 x”这句话可能没有太大意义。在理想情况下，我们掌握局部几何结构或组合结构方面的信息，从而可以着手进行分析。

一般而言，数据集 $\mathcal{U}$ 中点 x 的邻域 N 是 $\mathcal{U}$ 的一个大小适当的子集，其中包含 x。如果 $\mathcal{U}$ 是一个度量空间（或扩展伪度量空间），那么 N 具有直径 s（N），即对 $y \in N$而言的最大距离 d（x， y）。

包含关系 $N \subset \mathcal{U}$ 决定了 Vietoris-Rips复形V（N）$\subset V$（$\mathcal{U}$）的包含关系。

如果每个 $x \in \mathcal{U}$ 都有一个特定的邻域 N_x选择，如第 3 节所述，那么所有此类邻域的集合就决定了滤过复形的一个包含关系

$$N(\mathcal{U}) := \cup_{x\in\mathcal{U}}\ V(N_x) \subset V(\mathcal{U}),$$

这些复形通常按照距离进行滤过。我称 $N(\mathcal{U})$ 为由邻域族 $N = \{N_x\}$ 定义的邻域复形。

邻域复形 $V(N)$ 是全局 Vietoris-Rips 复形 $V(\mathcal{U})$ 关于覆盖 $V(N_x) \subset V(\mathcal{U})$ 的 Mapper 复形，如 [6] 所示。

邻域 N_x 中的每个元素 $y \neq x$ 都确定一条射线

$$\{x, y\} \subset N_x \subset \mathcal{U},$$

而这些射线的集合确定一个滤过子复形

$$R(N_x) = \vee_{y\neq x}\ V(\{x, y\}) \subset V(N_x).$$

取并集

$$R(\mathcal{U}) = \cup_{x\in\mathcal{U}}\ R(N_x) \subset V(\mathcal{U})$$

定义射线子复形 $R(\mathcal{U})$，它同时是 $V(\mathcal{U}$ 和 $N(\mathcal{U})$ 的子复形。

射线子复形 $R(\mathcal{U})$ 是一个经过筛选的（或加权的）图。

如果邻域 N_x 由 $\mathcal{U}$ 中各点的 k 个最近邻组成的集合构成，那么射线子复形 $R(\mathcal{U})$ 就是 k 个最近邻图，这是一个经过充分研究的对象。k 个最近邻图用于构建 [5], [3], [4] 的 UMAP 图。

滤过复形的包含映射

$$R(\mathcal{U}) \subset N(\mathcal{U}) \subset V(\mathcal{U})$$

诱导出满射

$$\pi_0 R_s(\mathcal{U}) \to \pi_0 N_s(\mathcal{U}) \to \pi_0 V_s(\mathcal{U})$$

对于距离参数 s，第3节和第4节将分析一些特殊情形。对于有界邻域中的有限 s，已有很好的比较结果，这正是第4节的主题。参见引理 14、引理15、引理16 和引理 20。

在这种情形下，有界邻域的邻域复形 $N_s(\mathcal{U})$ 在较小的距离尺度 s下，与全局 Vietoris-Rips 复形 $V_s(\mathcal{U})$ 具有相同的 1-骨架，因此，对于这类 s，邻域复形 $N_s(\mathcal{U})$ 可作为 $V_s(\mathcal{U})$进行聚类的良好近似。

在较大的距离尺度下，射线复形 $R_s(\mathcal{U})$ 的簇与邻域复形 $N_s(\mathcal{U})$ 的簇相同。因此，对于聚类而言，邻域复形 $N(\mathcal{U})$ 是射线复形 $R(\mathcal{U})$（一种类似 UMAP 的对象）与完整的 Vietoris-Rips 复形 $V(\mathcal{U})$ 之间的桥梁。

本文的基本思想和构造见第2节和第3节，并讨论了邻域与最近邻序列之间的关系。考虑到潜在应用（如第5节所述），我们通常假设 $\mathcal{U}$ 是一个扩展伪度量空间，或 ep-度量空间。关于 ep-度量空间的基本思想总结于第 1 节。

后续的结果和计算取决于邻域的选择，而这些选择会因具体示例的几何结构或组合结构而异。

第4、6和7节的定义和结果建立在朴素的示例（或思想实验）之上，用于启发和说明这些思想:

1）Gaia档案库 $\mathcal{U}$ 是一个数据库，收录了银河系中大约十亿颗恒星。该档案库的原始数据是一组扫描结果，由Gaia空间观测台航天器自2014年起采集。这些扫描以高分辨率返回小孔径内恒星的测光数据和光谱数据，因此，该档案库由局部数据汇集而成。经过反复观测和大量计算，确定了档案库中恒星相对于太阳的位置。

这些位置可以表示为一个函数 $p\colon\ \mathcal{U} \to \mathbb{R}^3$ 该函数能够唯一确定档案库 $\mathcal{U}$ 中的成员。位置函数 p 是一种降维。在 Mapper 构造的术语中，它是一个滤子函数 [6]。

根据观察，如果 x 是档案 $\mathcal{U}$ 中的一颗恒星，那么存在一个恒星邻域 $N_x \subset \mathcal{U}$，它靠近 x，且 N_x 中的元素数可计算。我们还可以要求 N_x 是一个有界邻域，即它具有有界半径 s（N_x），并且对于某个整数 k，其中至多包含 k 个元素。

这是一个明确的几何示例，与第4节介绍的方法密切相关。

2）对于某些数据集，存在一种图结构 Γ ，但没有明显的环境度量空间。

例如，在一个（足够短的）时间区间内，计算机账户之间的一组数据传输可以构造成稀疏有向加权图，如下文例30所示。每次传输所发送的字节数就是其权重。

这些图的顶点价数较低。对于每个账户 x，人们都知道它的所有传输e: $x \leftrightarrow y$ ，并据此构建一个可计算的账户邻域 $N_k(x)$，其中的账户与 x 之间最多相隔 k 个传输步数（或跳数）。

需要一种方法，为各种 $y \in N_k(x)$ 分配权重 $d(x,y)$。以某个账户 x为起点，可以认为与它进行“业务”往来最多的账户 y 离 x最近。 $N_k(x)$ 中离 x 最近的元素 y ，是根据 x 与 y之间所有有向边路径的权重总和 $\Sigma(x,y)$“反向”定义的。然后，距离 $d(x,y)$ 可定义为

$$d(x,y) = e^{-\Sigma(x,y)}$$

对于每个 $y \in N_k(x)$，在 x 和 y 之间。

根据邻域和权重数据，Healy-McInnes UMAP 机器会在图 Γ的顶点集合 Z 上生成一个全局 ep-度量 D ，其簇由有向集 $\pi_0 V(Z,D)$给出，或者等价地（定理 27），由源自各个邻域射线的有向集 $\pi_0 R(N)$给出$N_k(x)$。

归根结底，要点在于利用图结构，为稀疏加权有向图 Γ的所有顶点 x 找到可计算的加权邻域 $N_k(x)$。随后，这些局部结构会拼接起来，在 Γ的顶点全集上定义一个全局 ep-度量，并构造簇。

这些思想见于第6节。概括而言，它们同样适用于所有稀疏加权图。

这里有一个基本思想：UMAP 构造利用加权邻域提供的局部信息来创建全局空间层级结构并进行簇计算，无论是否存在环境度量。

这里展示的示例反复应用了这一观察。我们指定带权重的邻域，然后将这些邻域交给通用机制处理。

UMAP 构造的相关理论特性总结于第5节。该节以另一种方式呈现 UMAP 图：通过拼接射线来构造它，而不调用 UMAP 的大部分标准方法——参见定理 27。

3）第7节讨论了自然语言处理（NLP）中“连续词袋”模型里的单词邻域。掌握了这类邻域（并对权重作出恰当定义）后，我们再次使用

UMAP 方法在语料库的单词集合$\mathcal{L}$ 上构造 ep-度量空间结构。

第7节的方法可扩展到任意有限的数据元素字符串集合，其中可根据字符串内的接近程度定义局部度量。

在迄今展示的示例中，数据集的局部性质会随给定的几何结构或组合结构而变化。这些结构在一定程度上由期望的计算结果决定，并构成计算的起点。

最后，也可能会给出一个非常大的点云 $\mathcal{U}$，它具有 ep-度量空间结构，却没有其他信息，而我们希望据此近似（或发现）给定点 $x \in \mathcal{U}$的邻域 N_x。

在这种情况下，似乎别无选择，只能采用基于反复随机抽样的蛮力方法，目的是学习邻域的描述，或为 x学习一个“k-完备”邻域 N_x。第8节将介绍一种可能的方法。

本文的 k-完备邻域（见第2节和第4节）与点 x的 k近邻集合密切相关，但其优点在于定义唯一，因此在理论上更易于处理。当然，必须预先指定正整数 k。

目录

1 扩展伪度量空间 5 2 邻域 7 3 拓扑构造 10 4 有界邻域 12 4.1 k-有界邻域. 12 4.2 完备邻域.15 5 UMAP构造 16 6 加权有向图 20 7 词袋 22 8 抽样 24

1 扩展伪度量空间

扩展伪度量空间 (X, d)，此处称为 ep-度量空间，是一个集合 X，连同一个函数 $d\colon X \times X \to [0, \infty]$，满足以下条件:

1) $d(x,x)=0$,

2) $d(x,y)=d(y,x)$,

3) $d(x,z)\leq d(x,y)+d(y,z)$.

没有条件要求 $d(x,y)=0$ 蕴含 x 和 y 相同——“伪”这一形容词正是由此而来；这个结构之所以是“扩展”的，是因为我们允许距离为无穷大。

度量空间 (X,d) 是一种 ep-度量空间，其中 $d(x,y)=0$ 意味着$x=y$，且所有距离 $d(x,y)$ 都是有限的。

存在一个由 ep-度量空间构成的范畴 ep − met，其态射 $f:(X,d)\to(Y,d')$ 是函数 $f:X\to Y$，并且这些函数按照如下意义是非扩张的：对于所有 $x,y\in X$，都有 $d'(f(x),f(y))\leq d(x,y)$。

范畴 ep − met 是余完备的，也就是说，它具有所有小余极限。

实际上，余积 $\sqcup_i(X_i,d_i)$ 是不交并集合 $\sqcup_i X_i$，并配备由以下方式定义的 ep-度量 d

$$d(x,y)=\begin{cases} d_i(x,y) & \text{if } x,y\in X_i \text{ for some } i, \text{ and} \\ \infty & \text{otherwise.}\end{cases}$$

余等化子由一个商函数构造而成。设 (X,d) 是一个 ep-度量空间，且 $p\colon X\to Y$ 是一个满射函数。那么 Y 上存在一个 ep-度量 D，使得对于任意一对 $z,w\in Y$，

$$D(z,w)=\inf_P \sum d(x_i,y_i),$$

其中每条“路径” P 由点对 $\{x_i,y_i\}$，$i\leq n$ 构成，这些点对属于 X，且满足 $z=p(x_0)$、$w=p(y_n)$ 和 $p(y_i)=p(x_{i+1})$，其中 $i\leq n-1$。函数 p 定义了一个映射 $p\colon(X,d)\to(Y,D)$，它是 ep-度量空间之间具有商的泛性质的映射。

例1. 假设 (X,d) 和 (X,d') 是具有相同元素集合 X 的 ep-度量空间。那么，ep-度量空间范畴中的粘合（楔和）$(X,d)\vee(X,d')$ 是定义在 X 上的一种 ep-度量空间结构，满足

$$D(z,w)=\inf_P \sum D(x_i,x_{i+1}),$$

其中每条路径 P都是由 X 中的元素 $z=x_0,x_i,\ldots,x_n=w$组成的一个序列，并且

$$D(x_i,y_i)=\min\{d(x_i,x_{i+1}),d'(x_i,x_{i+1})\}.$$

每个有限 ep-度量空间 $\mathcal{U}$ 都有一族 Vietoris-Rips 复形 $V_s(\mathcal{U})$，它们由距离 s参数化。具体而言，$V_s(\mathcal{U})$ 是一个抽象单纯复形（或偏序集），其单形是 $\mathcal{U}$ 的有限子集 $\sigma=\{x_0,\ldots,x_k\}$，满足 $d(x_i,x_j)\leq s$。单纯形 σ 是一个 k-单纯形，其基数为$k+1$。

与标准情形一样，存在一个复形的递增族

$$V_s(\mathcal{U}) \subset V_t(\mathcal{U}),\ s \leq t,$$

其中 $\mathcal{U} = V_0(\mathcal{U})$ 是集合 $\mathcal{U}$ 上的离散复形。

极限对象 $V_\infty(\mathcal{U})$ 是一个以 $\mathcal{U}$ 为顶点的单纯形 $\Delta^{\mathcal{U}}$，但并非如此： $V_\infty(\mathcal{U})$ 是子对象 $V_s(\mathcal{U})$ 的不交并，且 s 有限。记

$$V_{<\infty}(\mathcal{U}) = \cup_{s<\infty}\ V_s(\mathcal{U}).$$

单纯集 $V_{<\infty}(\mathcal{U})$ 是有限个可缩分支的不交并。

2 邻域

设 (Z, d) 是一个有限 ep-度量空间，并且 $x \in Z$。

在以下所有情形中，

$$Z(x, s) = \{y \in Z \mid d(x, y) \leq s\}$$

是以 x 为中心、半径为 s 的 Z 闭球。

x 的邻域 N 是 Z 的子集 N，满足 $x \in N$ 且 $d(x, y) < \infty$。

邻域 N 从 Z 继承 ep-度量空间结构，并定义 Vietoris-Rips复形 $V(Z)$ 的滤过子复形 $V(N) \subset V(Z)$。

邻域 N 的半径 $s(N)$ 定义为

$$s(N) = \max_{y \in N}\ d(x, y).$$

根据假设， $s(N) < \infty$ 。

如果 $N = Z(x, s_N)$，则称邻域 N 是完备的。

若对于所有 $z \in Z - N$， $N$of x 都是最近邻的集合，且满足 $d(x, z) \geq s(N)$。如果 $N = \{x, x_1, \ldots, x_k\}$ 是一个最近邻集合（即其基数为 $k + 1$），那么 N 是一个 k-最近邻集合。

对于满足 $d(x, y) < \infty$ 的 x，其最近邻 y 可与一个最近邻邻域 $N = \{x, y\}$ 对应。这意味着，对于所有 $z \in Z - \{x\}$， $d(x, y) \leq d(x, z)$。一般而言，距离 $d(x, y)$ 可以为 0。

每个完备邻域 $N = Z(x, s_N)$ 都是 x 的一个最近邻集合，也是一个 n-最近邻集合，其中 $n = |N| - 1$。

引理 2. 假设 $N = \{x, x_1, \ldots, x_k\}$ 是 x 的一个最近邻集合，并且 x_i 按如下方式排序：

$$d(x, x_1) \leq d(x, x_2) \leq \cdots \leq d(x, x_k).$$

那么， x_i 是子集 $Z - \{x_1, \ldots, x_{i-1}\}$中 x 的最近邻。

证明。我们有

$$d(x, x_i) \leq d(x, x_{i+1}) \leq \cdots \leq d(x, x_k) \leq d(x, z)$$

对于所有位于 N之外的 z。由此可得，对于所有 $w \in Z - \{x_1, \ldots x_{i-1}\}$， $d(x, x_i)) \leq d(x, w)$。 □

引理 3. 假设邻域 N 是 x 的最近邻集合，并选取 $z \in X - N$ 使得对于所有$v \in Z - N$, $d(x, z) < \infty$，且 $d(x, z) \leq d(x, v)$。那么集合 $N \cup \{z\}$ 是 x 的最近邻集合。

证明. $N \cup \{z\}$ 的半径 s_z 为 $d(x, z)$。取 $v \in Z - (N \cup \{z\})$。那么$s(N) \leq d(x, v)$, 并且根据 $d(x, z)$ 的最小性， $d(x, z) \leq d(x, v)$。由此可得， $s(N \cup \{z\}) \leq d(x, v)$。 □

注 4. 归纳应用引理 3，可得到具有所有可能有限基数 $|N|$ 且满足 $|N| \leq |Z|$ 的 x 的最近邻邻域 N。

存在一个函数 $d_x \colon Z \to [0, \infty]$，满足 $d_x(y) = d(x, y)$。 x 的一个最近邻是满足 $d_x(z) < \infty$ 且 $d_x(z)$ 取最小值的元素 $z \in Z - \{x\}$。

对于这样的元素 z，记作 $s = d_x(z)$。那么 s 是像 $d_x(Z)$ 的最小有限值，而 $z \in Z_x(s)$，其中

$$Z_x(s) = d_x^{-1}(s)$$

是 d_x 在 s上的纤维（原像）。

引理 5. 假设 N 是 x的最近邻集合，并且 $\{s_1, \ldots, s_p\}$ 是像 $d_x(N)$ 中元素构成的集合，其中 $s_1 < \ldots s_p$。那么 $\{s_1, \ldots, s_p\}$ 是 $d_x(Z)$ 中最小有限元素构成的集合，并且

$$N = Z_x(s_1) \cup \cdots \cup Z_x(s_{p-1}) \sqcup F$$

其中 $F \subset Z_x(s_p)$。

证明。通过对 $|N|$ 作归纳，并利用引理 2 和引理 3，即可证明。 □

如果邻域 $N = \{x, x_1, \ldots, x_k\}$ 是 x的一个最近邻集合，且

$$d(x, x_1) \leq \cdots \leq d(x, x_k),$$

则称 $(x_1, x_2, \ldots, x_k)$ 是 x的一个由 k个最近邻构成的序列。

引理 6. 假设 $\{y_1, \ldots, y_k\}$ 是 $X - \{x\}$的一个由不同元素构成的集合，且

$$d(x, y_1) \leq d(x, y_2) \leq \cdots \leq d(x, y_k) < \infty.$$

如果（$x_1, \ldots, x_k$）是 x 的 k-最近邻序列，那么对于 $1 \leq i \leq k$， $d(x, x_i) \leq d(x, y_i)$。

证明。 $d(x, x_1) \leq d(x, y_1)$，因为 x_1 是最近邻。

设对于 $i \leq r$， $d(x, x_i) \leq d(x, y_i)$。那么

1) 如果 $d(x, x_r) < d(x, y_{r+1})$，那么根据最小性， $d(x, x_{r+1}) \leq d(x, y_{r+1})$。

2) 如果 $d(x, x_r) = d(x, y_{r+1})$，那么 y_{r+1} 是 $Z - \{x_1, \ldots, x_r\}$ 中 x 的最近邻，因此 $d(x, x_{r+1}) = d(x, y_{r+1})$。 □

推论 7. 设 $(x_1, \ldots, x_k)$ 和 $(y_1, \ldots, y_k)$ 是对于 x 的 k-最近邻序列。那么对于所有 i, $d(x, x_i) = d(x, y_i)$。

推论 8. 设 W 是有限 ep-度量空间，且包含映射 $Z \subset W$ 在子集 Z 上诱导出一个 ep-度量结构。假设 $x \in Z$。设 $(w_1, \ldots, w_k)$ 和 $(z_1, \ldots, z_k)$ 分别是 W 和 Z 中对于 x 的 k-最近邻序列。那么对于 $1 \leq i \leq k$， $d(x, w_i) \leq d(x, z_i)$。

引理 9. 假设 $(x_1, \ldots, x_k)$ 是 x 在 Z 中的最近邻序列，并且 $\{y_1, \ldots, y_k\}$ 是 Z 的一个由不同元素构成的序列，满足 $d(x, y_1) \leq \cdots \leq d(x, y_k) < \infty$。

如果对所有 i 都有 $d(x, x_i) = d(y, y_i)$，那么 $(y_1, \ldots, y_k)$ 是 x 的最近邻序列。

证明。对所有 z 都有 $d(x, y_1) = d(x, x_1) \leq d(x, z)$，因此 y_1 是 x 在 Z 中的最近邻。

归纳地，假设 $\{y_1, \ldots, y_i\}$ 是 x 的最近邻集合。

假设 $d(x, x_i) = d(x, x_{i+1})$。那么 $d(x, y_i) = d(x, y_{i+1})$，因此 $\{y_1, \ldots, y_{i+1}\}$ 是一个最近邻集合。

如果 $d(x, x_i) < d(x, x_{i+1})$，那么通过比较纤维 $Z_x(s)$，可得 $\{y_1, \ldots, y_i\} = \{x_1, \ldots, x_i\}$，因此 y_{i+1} 是 $Z - \{y_1, \ldots, y_i\}$ 中 x 的最近邻。 □

本节最后讨论 k-完备邻域。

距离函数 $d_x : Z \to [0, \infty]$ 的像具有如下形式：

$$\mathrm{Im}(d_x) = \{s_1, s_2, \ldots\},$$

其中，对于所有 i，不等式 $s_i < s_{i+1}$ 均为严格不等式。数据集 Z 是 d_x 的非空纤维的不交并：

$$Z = p_x^{-1}(s_1) \sqcup p_x^{-1}(s_2) \sqcup \cdots = Z_x(s_1) \sqcup Z_x(s_2) \sqcup \ldots.$$

对于每个 s_i，都存在唯一的 x 的完备邻域 $Z(x, s_i)$，满足

$$Z(x, s_i) = d_x^{-1}(s_1) \sqcup \cdots \sqcup d_x^{-1}(s_i).$$

x 的完备邻域构成一个有限的递增塔

$$Z(x, s_1) \subset Z(x, s_2) \subset Z(x, s_3) \subset \ldots$$

任何满足 $Z(x, s_i) \subsetneqq N$ 的完备邻域 N，其半径 $s_N > s_i$ 都必须严格更大。

假设 k 是一个正整数，且 $|Z| \geq k$。那么存在最小的数 i，使得 $|Z(x, s_i)| \geq k$。在这种情况下，邻域$Z(x, s_i)$ 是 k-完备的。

或者， x 的 k-完备邻域 N 是满足条件 $|N| \geq k$的最小完备邻域。

若满足 $|Z| \geq k$，则元素 $x \in Z$ 在 Z中有唯一的 k-完备邻域 N。 k-完备邻域 N 是定义良好的对象，而 x的 k-最近邻集合可能不止一个。

引理 10。设 Z_1、 $Z_2 \subset \mathcal{U}$，且 $x \in Z_i$。设 $N_i \subset Z_i$是 x 在 Z_i中的 k-完备邻域，并设 N 是 x 在 $Z = Z_1 \cup Z_2$中的 k-完备邻域。则 $N \subset N_1 \cup N_2$。

证明。集合 N 是 x 在 $Z_1 \cup Z_2$中的一个最近邻集合，而 $Z_i \cap N$ 是 x 在 Z_i中的一个最近邻集合。

实际上，若 $z \in Z_i$不在 $Z_i \cap N$中，则 z不在 N中，因此 $d(x, z) \geq s_N$，而 $s_N \geq s_{Z_i \cap N}$。

如果存在一个 $s < s_N$，使得 $|Z_i(x, s)| \geq k$，那么对于 $s < s_N$， $|Z(x, s)| \geq k$，因此 $N = Z(x, s_N)$ 不是 k-完备的。由此可得，对于 $s < s_N$， $|Z_i(x, s)| < k$，因此 $Z_i(x, s_N) \subset N_i$。

因此， $N \subset N_1 \cup N_2$，正如所声称的那样 □

引理 10 提供了一种方法，可用于近似非常大的数据集 $\mathcal{U}$ 中某个点 x 的 k-完备邻域。

实际上，如果 $Z_i \subset \mathcal{U}, 1 \leq i \leq p$ 是 $\mathcal{U}$ 的一个子集集合，且满足 $x \in Z_i$，并且如果 $N_i \subset Z_i$ 是 Z_i中 x 的一个 k-完备邻域，那么 x 在 $Z_1 \cup \cdots \cup Z_p$ 中的 k-完备邻域 N，就是较小对象 $N_1 \cup \cdots \cup N_p$中 x的 k-完备邻域。

3 拓扑构造

假设 (Z, d) 是一个有限的 ep-度量空间。

假设对每个 $x \in Z$ 给定一个邻域集合 N_x。回顾一下，邻域 N_x 的直径为 $s(N_x) < \infty$。

每个邻域 N_x 都确定 Vietoris-Rips复形 $V(Z)$ 的一个滤过子复形 $V(N_x) \subset V(Z)$

包含映射 $\{x, y\} \subset V(N_x)$、 $y \in N_x - \{x\}$ 诱导出滤过单纯复形映射

$$R(N_x) := \vee_y \Delta^1_{\geq s} \subset V(N_x) \subset V(Z). \tag{1}$$

Δ^1 的副本由权重为 s 的射线 $\{x, y\}$ 定义。

注11。更准确地说，若射线 $\{x, y\}$ 的权重为 $t = d(x, y)$，则 $R(N_x)$ 对应的 1-单形是滤过单形 $\Delta^1_{\geq t}$，满足

$$(\Delta^1_{\geq t})_s = \begin{cases} \emptyset & \text{if } s < t, \text{ and} \\ \Delta^1 & \text{if } s \geq t. \end{cases}$$

有时，更好的说法是， $R(N_x)$ 被与权重为 s 的射线 $\{x, y\}$ 相对应的单形 $\Delta^1_{\geq s}$ 覆盖。这仅仅反映了这样一个事实：显然的映射

$$\sqcup_y \ \Delta^1_{\geq s} \to \vee_y \ \Delta^1_{\geq s} = R(N_x)$$

是滤过复形的一个满同态。

完整并集

$$R(N) = \cup_x \ R(N_x) \subset V(Z)$$

是针对邻域集合 $N = \{N_x\}$ 的 $V(Z)$ 的射线子复形。

射线子复形 $R(N)$ 是一个滤过（或加权）图。若邻域 N_x 由 k-最近邻构成，则 $R(N)$ 是 k-最近邻（kNN）图。

邻域 N_x 生成一个抽象单纯复形 $V(N) \subset V(Z)$，其单形是各个邻域N_x的子集 $\sigma \subset N_x$ 。所得的滤过单纯复形可写为

$$V(N) = \cup_x \ V(N_x) \subset V(Z).$$

$V(Z)$ 的子复形 $V(N)$ 称为邻域复形。

包含映射 $R(N_x) \subset V(N_x)$ 诱导出一个包含映射 $R(N) \subset V(N)$，因此我们有包含映射

$$R(N) \subset V(N) \subset V(Z) \tag{2}$$

滤过复形之间的包含映射，以及相应的包含映射

$$R_s(N) \subset V_s(N) \subset V_s(Z) \tag{3}$$

各个滤过阶段之间。

诱导的函数

$$\pi_0 R_s(N) \to \pi_0 V_s(N) \to \pi_0 V_s(Z)$$

在道路连通分支（或簇）之间的函数对所有参数 t 都是满射，因为所有复形具有相同的顶点集合，即 Z。

注12. 邻域复形 $V(N) = \cup_x V(N_x)$ 被子复形 $V(N_x)$ 覆盖，其含义是存在一个满射

$$\bigsqcup_x V(N_x) \to V(N).$$

˘ 这个覆盖有一个相关的切赫分解，并且存在一个自然的余等化子

$$\bigsqcup_{x,y} V_s(N_x) \cap V_s(N_y) \rightrightarrows \bigsqcup_x V_s(N_x) \to V_s(N)$$

在单纯集中，其中 s 是距离参数。道路分支函子保余极限，因此存在一个余等化子

$$\coprod_{x,y} \pi_0(V_s(N_x) \cap V_s(N_y)) \rightrightarrows \coprod_{x} \pi_0 V_s(N_x) \to \pi_0 V_s(N)$$

在集合图表中，或簇中。

有向集 $\pi_0 V_s(N)$ 是由映射器构造为用邻域族 $\{N_x\}$ [6] 覆盖 Z 而得到的簇对象。

引理 13. 假设 $x, y \in Z$。在 $R(N)$ 中，存在一条从 x 到 y 的路径，当且仅当存在一个元素序列

$$x = x_0, x_1, \ldots, x_r = y$$

以及 x_i 的邻域 N_{x_i}，使得对于所有 i，$N_{x_i} \cap N_{x_{i+1}} \neq \emptyset$。

证明。假设

$$x = z_0, \ldots, z_p = y$$

是一个点序列，使得对于 x_i 和 x_{i+1} 的邻域 N_{x_i} 和 $N_{x_{i+1}}$，分别有 $x_{i+1} \in N_{x_i}$ 或 $x_i \in N_{x_{i+1}}$。若 $x_{i+1} \in N_{x_i}$，则 $N_{x_i} \cap N_{x_{i+1}} \neq \emptyset$。同理，若 $x_i \in N_{x_{i+1}}$，则 $N_{x_i} \cap N_{x_{i+1}} \neq \emptyset$。

反过来，假设 $v \in N_{x_i} \cap N_{x_{i+1}}$。那么，$N_{x_i}$ 中有一条边 $x_i \to v$，且 $N_{x_{i+1}}$ 中有一条边 $x_{i+1} \to v$，因此存在一条路径

$$x_i \to v \leftarrow x_{i+1}$$

穿过若干邻域。 □

根据定义，射线复形 $R(N)$ 是 $V(Z)$ 的一个滤过子复形。子复形 $R_s(N) \subset V_s(Z)$ 由满足 $d(x, y) \leq s$ 的射线 $\{x, y\}$ 生成。

我们有如下引理 13 的类似结果：

引理 14。假设 $x, y \in Z$。对于每个参数值 s，在 $R_s(N)$ 中存在一条从 x 到 y 的路径，当且仅当存在一个元素序列

$$x = x_0, x_1, \ldots, x_r = y$$

以及 x_i 的邻域 N_{x_i}，使得对于所有 i，$(N_{x_i})_s \cap (N_{x_{i+1}})_s \neq \emptyset$。

4 有界邻域

4.1 k-有界邻域

在一些例子中（例如星图），邻域 N的 x具有有界基数和半径是很自然的：$|N| \leq k+1$对于某个 k和 s（N） $\leq S$，其中 k 和 S 都是固定的。

从这个角度来看，对于固定的 x， k-有界邻域 N的 x是 Z（x， S）的子集，其中包含 x且至多有 $k+1$个元素。同样， Z（x， S）是 Z中半径为 S、以 x为中心的球。

我们在此后假设 $k \geq 1$。

点 x 可以有多个 k-有界邻域。 x 的 k-有界邻域按包含关系排序，且这个族有极大元素。我们有以下结论：

1) 极大 k-有界邻域 $N \subset Z(x, S)$ 的基数要么为 $k+1$，要么满足 $N = Z(x, S)$。

2) 所有满足 $s(N) \leq S$ 的 k-最近邻集合 N 都是极大的。

3) 如果 $N = \{x\}$ 是极大的，那么对于参数 S， x 是一个孤立点。

相应的邻域复形 $V(k-N)$ 是 $V(Z)$ 的滤过子复形，由所有 x 的 k-有界邻域 N 对应的子对象 $V(N)$ 生成，而 $R(k-N)$ 是相应的射线子复形。与 (2) 中一样，我们有一个包含映射序列

$$R(k-N) \subset V(k-N) \subset V(Z).$$

如果 $t \leq S$且 $\sigma = \{x_0, \ldots, x_n\}$是 $V(Z)_t$的一个 n-单纯形，且满足 $n \leq k$，那么 σ是 x_0的一个 k-有界邻域。实际上， σ至多有 $k+1$个元素，其到 x_0的最大距离为 $t \leq S$。由此可得，对于 $n \leq k$， $V_t(k-N)_n = V_t(Z)_n$，或者 $\mathrm{sk}_k V_t(k-N) = \mathrm{sk}_k V_t(Z)$。特别地，由于 $k \geq 1$， $\mathrm{sk}_1 V_t(k-N) = \mathrm{sk}_1 V_t(Z)$，因此单纯集 $V_t(k-N)$ 和 $V_t(Z)$ 具有相同的路径连通分支。

我们已经证明了以下结论：

引理15。假设 $t \leq S$，并按照上述方式由 k-有界邻域构造邻域复形 $V(k-N)$。那么该函数

$$\pi_0 V_t(k-N) \to \pi_0 V_t(Z)$$

是一个双射。

假设 $t \geq S$，且 $\{x, y\}$是 $V_t(k-N)$ 的一个 1-单形。那么，对于某个 z 的 k-有界邻域 N，$\{x, y\} \subset N$。此外， $d(z,x) \leq s(N)$ 且 $d(z,y) \leq s(N)$，因此 $d(z,x)\ d(z,y) \leq s(N) \leq S \leq t$。由此可知存在一条路径

$$x \leftarrow z \rightarrow y$$

在 $R_t(k-N)$ 中，因此该函数

$$\pi_0 R_t(k-N) \to \pi_0 V_t(k-N)$$

是一个双射。

我们已经证明

引理 16. 假设 $t \geq S$。那么诱导出的函数

$$\pi_0 R_t(k-N) \to \pi_0 V_t(k-N)$$

是一个双射。

记

$$R(k-N) = \cup_t \ R_t(k-N).$$

那么映射

$$\pi_0 R_t(k-N) \to \pi_0 R(k-N)$$

对于 $t \geq S$，它是一个双射，因为在该范围内，$R_t(k-N) = R(k-N)$。

因此，我们得到如下结论：

推论 17. 函数

$$\pi_0 R(k-N) \leftarrow \pi_0 R_t(k-N) \to \pi_0 V_t(k-N)$$

对于所有 $t \geq S$ 都是双射。

记

$$N(x) = \cup_N \ V(N)$$

在 $V(Z)$ 中，其中的并是对所有 k-有界邻域 N of$x$ 取的。令 $R(x) \subset N(x)$ 为相应的射线子复形。我们有包含映射

$$VR(x) \subset VN(x) \subset V(Z(x,S)).$$

假设 $t \leq S$and $n+1 \leq k$。假设 $\sigma = \{x_0, \ldots, x_n\}$ 是 $V(Z(x,S))_t$ 的一个非退化 n-单纯形。记

$$\sigma_x = \{x, x_0, \ldots, x_n\}.$$

于是 $|\sigma_x| \leq k+1$，因此 σ_x 是 x 的一个 k-有界邻域，所以 $\sigma = d_0\sigma_x$ 属于复合映射的像

$$V(\sigma_x)_t \to VN(x)_t \to V(Z(x,S))_t.$$

由此可得，对于 $t \leq S$，$\mathrm{sk}_{k-1} V_t N(x) = \mathrm{sk}_{k-1} V_t(Z(x,S))$。特别地，映射

$$\pi_0 V_t N(x) = \pi_0 V_t(Z(x,S))$$

当 $k \geq 2$ 和 $t \leq S$时，是一个双射。

假设 $t \geq S$，且 y和 z是 $N(x)$ 的顶点。那么 $d(x,y)\ d(x,z) \leq S \leq t$，由此可知映射

$$* = \pi_0 R(x)_t \to \pi_0 N(x)_t$$

是一个双射。

每个 $y \in Z(x,S)$ 都是一个 k-有界邻域 $\{x,y\}$ 的成员，因为$k \geq 1$。由此可知映射

$$\pi_0 R(x)_t \to \pi_0 N(x)_t \to \pi_0 V_t Z(x,S)$$

是满射。

我们已经证明：

引理 18. 设复形 $N(x)$ 和 $R(x)$ 如上所定义。设 $k \geq 2$。那么我们有如下结论:

1) If $t \leq S$ then the map $\pi_0 N(x)_t \to \pi_0 V_t(Z(x,s))$ is a bijection.

2）如果 $t \geq S$，则映射

$$* = \pi_0 R(x)_t \to \pi_0 N(x)_t \to \pi_0 V(Z(x,S))_t$$

是双射。

4.2 完备邻域

设 Z 是一个有限 ep-度量空间，并且每个 $x \in Z$ 都具有一个固定的完备邻域 $N_x = Z（x, r_x）$。构造相应的滤过复形

$$R(N) \subset V(N) \subset V(Z),$$

对于 Z 以及邻域系 $N = \{N_x\}$。

例 19。设 $S > 0$ 是一个固定的距离参数，且 $k > 1$ 是一个固定的整数。

称邻域 N_x $x \in Z$ 为完备的 k-有界邻域，若 N_x 具有如下形式

$$N_x = Z(x, s_x) \cap Z(x, S),$$

其中 $Z(x, s_x)$ 是 x 的唯一 k-完备邻域（见第 2 节）。

有两种可能： $N_x = Z(x, s_x)$，此时 N_x 是 k-完备的；或者 $N_x = Z(x, S)$ 且 $|Z(x, S)| < k$。无论哪种情况，邻域 N_x 都是唯一确定的，并且是完备的。

采用完备的 k-有界邻域，为恒星图表示例提供了不同的视角。对于固定的（且适当的）距离 S 和正整数 k，球状星团中恒星 x 的完备 k-有界邻域 N_x 将是半径较小的 k-完备邻域；而外侧旋臂中的恒星则更可能具有基数较小的邻域 N_x 。

引理 20. 假设 Z 是一个有限 ep-度量空间，并且每个 $x \in Z$ 都有一个固定的完备邻域 $N_x = Z(x, r_x)$。

1）假设对于所有 x， $t \leq r_x$ 。那么这些函数

$$\pi_0 R_t(N) \to \pi_0 V_t(N) \to \pi_0 V_t(Z)$$

都是双射。

2）假设对于所有 x， $t \geq r_x$ 。那么这个映射

$$\pi_0 R_t(N) \to \pi_0 V_t(N)$$

是一个双射。

3）假设对所有 x 都有 $t \geq S \geq r_x$。则映射

$$\pi_0 R_S(N) \to \pi_0 R_t(N)$$

是一个双射。

证明。对于 1），假设 $\{x, y\}$ 是一个长度为 t的 1-单形。那么 $y \in N_x$，因为 $t \leq r_x$，且 $\{x, y\}$ 是 N_x的一条射线。由此可得 1-骨架相等：

$$\mathrm{sk}_1 R_t(N) = \mathrm{sk}_1 V_t(N) = \mathrm{sk}_1 V_t(Z),$$

结论随即成立。

对于陈述 2），假设 $\{x, y\}$ 是 $V_t(N_z) \subset V_t(N)$的一个 1-单形。那么由于 $t \geq r_z$，$V_t(N_z)$ 中存在 1- 单形 $x \leftarrow z \to y$。这对所有 z 都成立，因此函数

$$\pi_0 R_t(N) \to \pi_0 V_t(N)$$

是一个双射。

为证明陈述 3），$R(N)$的每条射线（1-单形）$\{x, y\}$ 的长度均为 $\leq S$，因此 $R_S(N) = R_t(N)$。 □

推论 21. 假设对于所有 $x \in Z$，都有 $t \geq S \geq r_x$。那么，邻域复形 $V_S(N) \subset V_t(N)$ 的包含映射诱导出一个双射

$$\pi_0 V_S(N) \to \pi_0 V_t(N).$$

证明. 该推论由引理 20 的陈述 2) 和 3) 得出 □

注 22. 假设 Z' 是由满足 $Z(x, r_x) = \{x\}$ 的元素 $x \in Z$ 构成的子集，并令 $Z'' = Z - Z'$。那么

1) $V_t(Z) = Z' \sqcup V_t(Z'')$,
2) $V(N)_t = Z' \sqcup V(N'')_t$,
3) $R(N)_t = Z' \sqcup R(N'')_t$,

对于 $t \leq r_z$，所有 z，其中 Z' 是一个离散集。这里，

$$Z'' = \cup_{y \in Z''} Z(y, r_y),$$

并且 N'' 是 $y \in Z''$ 的邻域系 $Z(y, n_y)$。

5 UMAP 的构建

首先，对于数据集 Z 的每个顶点 x，取一个邻域 N_x，并为每个 $y \in N_x - \{x\}$ 配备正权重 $d(x, y)$。对于这样的 y，子集 $\{x, y\}$ 被称为射线。

权重 $d(x,y)$ 在集合 $\{x,y\}$ 上定义了一个 ep-度量空间结构。构造 ep-度量空间

$$Z_x = \vee_{y\in N(x)-\{x\}}\ \{x,y\},$$

对于每个 $x \in Z$，来自射线 $\{x,y\}$。通过设定以下内容，可将此结构扩展为顶点全集 Z of Γ 上的 ep-度量空间结构 (Z, D_x):

$$(Z, D_x) = (\sqcup_{z\in Z-Z_x}\ \{z\}) \sqcup Z_x$$

在 ep-度量空间中。

该 ep-度量空间

$$(Z, D) = \vee_{x\in Z}\ (Z, D_x)$$

以及 UMAP 复形

$$V(Z, N) = \vee_{x\in Z}\ V(Z, D_x)$$

分别在 ep-度量空间和滤过复形中沿顶点（ Z 的元素）进行粘合而成。

对于这些 ep-度量空间构造，了解 ep-度量空间范畴是余完备的这一点至关重要——参见第 1 节。

以下关于道路连通分支的切除陈述是 [3] 的引理 2:

定理 23。典范映射 $V(Z,N) \to V(Z,D)$ 诱导同构

$$\pi_0 V(Z,N)_s \xrightarrow{\cong} \pi_0 V(Z,D)_s$$

对于 s 有限。

定理 23 的证明思路是注意到，(Z,D) 中的距离是通过邻域 N_x 中的路径计算得出的。

我们将需要以下局部计算:

引理 24。假设 $y \in N(x) - \{x\}$ 定义了射线 $\{x,y\}$。那么

$$d(x,y) = D_x(x,y).$$

在 Z_x 中。

证明。数 $D_x(x,y)$ 是所有和 $\sum_j d(x_j, x_{j+1})$ 中的最小值，其中路径

$$P:\ x = x_0, x_1, \ldots, x_p = y$$

穿过 Z_x 中的射线。射线 $\{x_{p-1}, y\}$ 必须是射线 $\{x,y\}$，因此

$$D_x(x,y) \geq \sum_j d(x_j, x_{j+1}) \geq d(x,y).$$

子对象 $\{x,y\}$ 是一条射线，因此 $\{x,y\}$ 是一条路径，并且 $D_x(x,y) \leq d(x,y)$。 □

每个 $y \in N_x - \{x\}$ 都确定一个滤过复形的包含关系

$$V(\{x,y\}) \subset V(Z_x, D_x) \subset V(Z, D).$$

单纯集 $V_t(\{x,y\}$ 由顶点 $\{x,y\}$ 组成，适用于 $t < d_x(x,y)$，并且其 1-单形为

$$\{x\} \subset \{x,y\} \supset \{y\}$$

适用于 $t \geq d_x(x,y)$。

注 25. 回想一下， $V(\{x,y\})$ 是一个经过滤的 1-单形的重心细分，该 1-单形可通过在集合 $\{x,y\}$ 上施加全序来定义。

设 $\{x,u\}$ 是 N_x 中的一条射线，且 $\{y,v\}$ 是 N_y的一条射线，并考虑复合单态射

$$V(\{x,u\}) \subset V(Z_x) \subset V(Z), \quad V(\{y,v\}) \subset V(Z_y) \subset V(Z).$$

设 $d_x(x,u) \leq d_y(y,v)$。

一般而言， $V(X) = BP(X)$，其中 $P(X)$ 是生成单形构成的偏序集。在当前情况下，有一个拉回图

$$\begin{array}{ccc} BP(\{x,u\} \cap \{y,v\}) & \longrightarrow & BP(\{y,v\}) \\ \downarrow & & \downarrow \\ BP(\{x,u\}) & \xrightarrow{\geq} & BP(Z) \end{array}$$

交集 $\{x,u\} \cap \{y,v\}$ 至多是一个含两个元素的集合。如果 $\{x,u\} \cap \{y,v\} = \emptyset$，则拉回为空；如果 $\{x,u\} \cap \{y,v\}$ 是一个点，则拉回也是一个点。

如果 $\{x,u\} \cap \{y,v\}$ 是一个含两个元素的集合，那么 $\{x,u\} = \{y,v\}$，并且存在一个交换图

$$\begin{array}{ccc} BP(\{x,u\}) & \xleftarrow{\theta} & BP(\{y,v\} \\ & \searrow \quad \swarrow & \\ & BP(Z) & \end{array}$$

其中 θ “降低权重”。因此，在这种情况下，存在一个拉回

$$\begin{array}{ccc} BP(\{y,v\}) & \xrightarrow{1} & BP(\{y,v\}) \\ \theta \downarrow & & \downarrow \\ BP(\{x,u\}) & \longrightarrow & BP(Z) \end{array} \tag{4}$$

射线复形 $R(N_x) \subset V(Z_x)$ 是射线楔

$$R(N_x) = \vee_{y \in N_x - \{x\}} V(\{x,y\}).$$

滤过复形单态

$$\phi_x : R(N_x) \to V(Z_x, D_x) \to V(Z, D),$$

共同定义一个单态

$$\phi : R(N) = \cup_{x \in Z}\ R(N_x) \subset V(Z),$$

并且我们称并集 $R(N)$ 为 $V(Z)$ 的射线子复形。射线复形 $R(N)$ 是一个加权图。

射线复形 $R(N)$ 是滤过子复形$V(\{x, y\})$ 的并（或由其覆盖），这些子复形由射线 $\{x, y\}$ 及其权重 $d(x, y)$ 定义。交（拉回）

$$\begin{array}{ccc} BP(\{y,v\} \cap \{y,v\}) & \longrightarrow & BP(\{y,v\}) \\ \downarrow & & \downarrow \\ BP(\{x,u\}) & \longrightarrow & R(N) \end{array}$$

如上所述，在 $V(Z)$ 中构造，因为 $R(N) \subset V(Z)$ 是一个单态。因此，射线复形 $R(N)$ 是射线的并，交集中的权重可能会进行调整，如拉回图 (4) 所示。

注 26. 对射线复形 $R(N)$ 的当前描述独立于空间 (Z, D) 中的距离。它推广了第 3 节中出现的射线复形描述；该描述使用固定的环境 ep- 度量。

最后，还有一个切除结果，使得 $R(N)$ 成为 UMAP 图的候选对象，如下所示：

定理 27。滤过复形映射 ϕ：$R(N) \subset V(Z, D)$ 诱导同构

$$\phi_* : \pi_0 R_s(N) \xrightarrow{\cong} \pi_0 V_s(Z, D)$$

对所有 $s \geq 0$ 均成立。

证明。该证明与定理 23 的证明类似。

映射 ϕ 在顶点上是恒等映射，因此函数 ϕ_* 是满射。

假设 $V(Z, D)_s$ 中存在一个 1-单形 $\{z, w\}$。那么存在一条路径

$$z = x_0, x_1, \ldots, x_p = w$$

穿过射线 $\{x_i, x_{i+1}\}$，使得

$$\sum_{i=0}^{p} d(x_i, x_{i+1}) \leq s.$$

根据引理 24。但随后对于所有 i，$d(x_i, x_{i+1}) \leq s$，因此 z 和 w 位于单纯集 $R(N)_s$ 的同一道路连通分支中。

由此可知，函数 ϕ_* 是单射。 □

6 加权有向图

加权有向图 Γ 由边集合 e 构成： $x \to y$，其中每条边 e 都有一个权重 $w(e)$ > 0。在当前讨论中， Γ 的顶点是边的面。我用 Z 表示 Γ的顶点集合。

简单的例子是定向单纯复形K的1-骨架$_1\, K$，其每个1-单形 e： $x \to y$都带有权重 $w(e) = 1$，并且每个顶点都是某个非退化1-单形 e的面。

如果加权有向图 Γ 的所有顶点 x 的价数都很低，就称它 Γ 为稀疏图。这意味着 Γ 的每个顶点都位于少量（即数量可计算的）边的边界上。

设 x 是传输图 Γ的一个顶点。从x 到 Γ中另一个顶点 y 的路径 P： $x \dashrightarrow y$ 是一个边串

$$P: \ x = y_0 \xrightarrow{e_1} y_1 \xrightarrow{e_2} \dots \xrightarrow{e_p} y_p = y. \tag{5}$$

称整数 p 为路径 P的长度 $\ell(P)$。

注28. 图 Γ 中所有路径 P： $x \dashrightarrow y$ 构成一个加权图 $P(\Gamma)$，其顶点与图 Γ 相同。路径P： $x \dashrightarrow y$ 和 Q： $y \dashrightarrow z$ 可以复合：将 P 与 Q连接起来，就定义了一条路径 $P \circ Q$： $x \dashrightarrow z$。因此， $P(\Gamma)$具有更多结构： $P(\Gamma)$是图 Γ 上的自由范畴。

路径图 $P(\Gamma)$ 通常并非稀疏。

路径 P 的权重 $w(P)$ 可定义为

$$w(P) = \min_i \{w(e_i)\}. \tag{6}$$

注29. 路径权重的定义在某种程度上是任意的，并取决于具体应用。式 (6) 中的赋值是受数据传输图的启发，这些图将在下文讨论。另一种做法是，令

$$w(P) = \sum_i w(e_i).$$

固定一个正整数 k。

邻域 $N_k(x)$ 是所有顶点 y 的集合，这些顶点出现在路径中

$$Q: \ z_0 \to z_1 \to \dots \to z_p,$$

长度为 $p = \ell(Q) \leq k$，使得 $x = z_i$ 对某个 i 成立

我们为邻域 $N_k(x)$ 中的所有 y 指定权重（或距离） $d(x, y)$。对于 $y \in N_k(x)$，将权重和 $\Sigma(x, y)$ 定义为

$$\Sigma(x,y) = \Big(\sum_{P: x \dashrightarrow y, \ell(P) \leq k} w(P)\Big) + \Big(\sum_{Q: y \dashrightarrow x, \ell(P) \leq k} w(Q)\Big).$$

在交易图中，权重和 $\Sigma(x,y)$ 表示 x 与 y 之间所有交易的总价值。如果 $\Sigma(x,y)$ 的值较大，那么 x 与 y 之间的业务就更多，这些对象在某种意义上应该更接近。为了表达这种关系，使用香农信息函数来定义距离

$$d(x,y) = e^{-\Sigma(x,y)} \tag{7}$$

对于 $y \in N_k(x)$。

当然也可以采用其他方法，为图 $N_k(x)$ 的顶点 y 定义距离 $d(x,y)$。

最终，对于稀疏有向图 Γ 的每个顶点 x，我们得到一个可计算的顶点邻域 $N_k(x)$，其距离（权重） $d(x,y)$ 为$y \in N_k(x) - \{x\}$。

这些就是 UMAP 构造的输入，相关内容将在第4节中介绍。

例30（数据传输）。一次数据传输 e： 从计算机账户 x 向另一个账户 y 传输 $x \to y$，其权重为 $w(e)$，即传输的字节数。该传输 e 还具有源时间戳和目标时间戳，分别为 $s(e)$ 和 $t(e)$，其中 $s(e) < t(e)$。因此（暂时地），数据传输图 Γ 的边为 $e: (x, s(e)) \to (y, t(e))$，其中 $x \neq y$；其顶点由成对元素 (x,t) 组成，其中 x 是计算机账户， t 是某条边的源时间戳或目标时间戳。

对于固定账户 x，可能存在多个顶点（x、t）。假设 $t_0 < t_2 < \cdots < t_p$ 是固定账户 x 的时间戳。将列表 $\{t_0, \ldots, t_p\}$ 称为账户 x 的时间戳单纯形。

对于 $x \neq y$，传输图 Γ 的一条边 $E: (x,s) \to (y,t)$ 由一次传输 $e: (x, s(e)) \to (y, t(e))$，以及关系 $s \leq s(e)$ 和 $t(e) \leq t$ 组成；这些关系分别属于账户 x 和 y 的时间戳单形。权重 $w(E)$ 是传输 e 的权重 $w(e)$。Γ 的顶点集合 Z 由所有账户 x 与传输时间戳 s 的配对 (x,s) 组成。

如果所有时间戳都位于足够小的区间内，那么传输图 Γ 就是稀疏的。

这个例子引出了上文所述的路径权重和邻域内距离的定义。

具体来说，一条路径

$$P: \ (x,s) = x_0 \xrightarrow{E_1} x_1 \xrightarrow{E_2} \ldots \xrightarrow{E_p} x_p = (y,t)$$

在 Γ 中由边组成

$$E_i : x_i = (x_i, s_i) \to (x_{i+1}, s_{i+1}) = x_{i+1}$$

具有 $x_i \neq x_{i+1}$，且每条这样的边的权重为 $w(E_i)$。

路径 P 的权重 $w(P)$ 定义为

$$w(P) = \min_i \{w(E_i)\},$$

如式（6）所示）。权重 $w(P)$ 表示沿路径 P从 (x,s) 传输到 (y,t) 的最大数据量。

取一个正整数 k 和传输图 Γ中的一个元素 $x=(x,s)$，并将邻域 $N_k(x)$定义为穿过 x 且长度至多为 k的路径上的顶点。

$y\in N_k(x)$ 的权重和 $\Sigma(x,y)$ 定义为

$$\Sigma(x,y)=(\sum_{P:x\dashrightarrow y,\ell(P)\le k} w(P))+(\sum_{Q:y\dashrightarrow x,\ell(P)\le k} w(Q)),$$

射线 $\{x,y\}$ 的权重 $d(x,y)$ 具有如下形式

$$d(x,y)=e^{-\Sigma(x,y)}.$$

注 31（无向图）。图 Γ的有向结构是上述示例中的核心特征。类似的局部到整体方法同样适用于为无向图构造 ep-度量空间和 UMAP 复形。

设 Ω 是一个稀疏加权图，其边e 的权重为 $w(e)$。假设 Ω 的顶点是其边的面。

设 x 是 Ω 的一个顶点。若存在一条路径（path）或边串，则称 $y\in N_k[x]$

$$P:\ x=x_0\overset{e_1}{\leftrightarrow}x_1\overset{e_2}{\leftrightarrow}\dots\overset{e_p}{\leftrightarrow}x_p=y$$

与 $p\le k$。

同样，这里也有多种选择，但将路径 P 的权重 $w(P)$ 定义为：$x\leftrightarrow y$由

$$w(P)=\min_i\{w(e_i)\}.$$

固定一个顶点 x 和一个正整数 k。定义 $N_k(x)$ 为 Ω 中所有满足以下条件的顶点所构成的集合：它们位于长度至多为 k 的路径上，且该路径经过 x，其中路径长度为 $\ell(P)$。记为

$$\Sigma(x,y)=\sum_{P:x\dashrightarrow y,\ell(P)\le k} w(P),$$

并令

$$d(x,y)=e^{-\Sigma(x,y)}$$

对于 $y\in N_k(x)$。

利用权重 $d(x,y)$，可以在邻域$N_k(x)$上构造一个 ep-度量。这些 ep-度量拼接起来，便在 Ω 的顶点全集 Z上给出一个 ep-度量。

7 个词袋

在自然语言处理的“词袋”模型中（例如参见[1]），首先从一组 $C=\{C_1,\dots,C_N\}$ 文档 C_i开始，其中

每个 $C_i = (t_{i,\,1},\ \ldots\ldots,\ t_{i,\,M_i})$都是由词元（即单词、短语等）组成的序列，词元可以重复。这个序列 C 就是语料库。

序列 C_i 是一个函数

$$C_i : \underline{M}_i \to \mathcal{T},$$

其中，$\mathcal{T}$是所有 C_i中不同词元组成的集合，而$\underline{M}_i = \{1, 2, \ldots, M_i\}$。序列 C_i 中可能有重复项，因此函数 C_i 通常不是单射。

通常的做法是将一些词元合并（依据词根或其他方式），从而构成满射映射 $\ell:\ \mathcal{T} \to \mathcal{L}$。集合 $\mathcal{L}$ 是词汇表，其中的元素称为单词。

用 p 表示复合函数

$$p : \sqcup_i\ \underline{M}_i \xrightarrow{C} \mathcal{T} \xrightarrow{\ell} \mathcal{L},$$

令 p_i: $\underline{M}_i \to \mathcal{L}$为 p在直和项 $\underline{M}_i$上的限制。

我们假设集合 $\mathcal{T}$中没有共有词元（“停用词”），也没有罕见词元，尽管这些词元的判定方式不尽相同。这意味着函数 p 的纤维 $p^{-1}(w)$既不会太大，也不会太小，因此在计算上可处理。函数 p 及其纤维是本讨论关注的对象。

纤维 $p^{-1}(w)$ 是语料库 C 中单词 $w \in \mathcal{L}$ 的实例。

注 32. 更一般地，我们可以考虑覆盖集合 Z 的函数 $p_i : \underline{M}_i \to Z$，也就是说，合并后的函数

$$p : \sqcup_i\ \underline{M}_i \to Z$$

是满射。这里，p 在直和项 $\underline{M}_i$ 上的限制是 p_i。假设对于 $z \in Z$，纤维 $p_i^{-1}(z)$ 在计算上可处理（或者按照下一节的意义是易处理的），函数 $\{p_i\}$ 的集合同样如此。

在基数 $\underline{M}_i$ 和函数集合 p_i 满足大小假设的前提下，以下讨论可应用于这种情形。

在接下来的讨论中，我们甚至可以用度量空间替换集合 $\underline{M}_i$。

用 $\mathcal{L}_i$ 表示受限制函数的像

$$p_i = \ell_i : \underline{M}_i \xrightarrow{C_i} \mathcal{T} \xrightarrow{\ell} \mathcal{L}.$$

复合映射 p_i 限制为一个满射 $p_i : \underline{M}_i \to \mathcal{L}_i$，并且存在如下函数交换图

$$\begin{array}{ccc} \underline{M}_i & \xrightarrow{p_i} & \mathcal{L}_i \\ \downarrow & & \downarrow \\ \sqcup_i\ \underline{M}_i & \xrightarrow[p]{} & \mathcal{L} \end{array}$$

其中垂直映射为包含映射。

每个集合 $\underline{M}_i$ 都有一个度量 d，满足 $d(x,y) = |y-x|$。

设 r 为正整数。固定一个单词 $v \in \mathcal{L}$，并设$p_i^{-1}(w)_{\leq r}$ 是所有元素 $y \in p_i^{-1}(w)$ 的集合，使得对于某个 $x \in p_i^{-1}(v)$，有 $d(x,y) \leq r$。于是我们有

$$p_i^{-1}(w)_{\leq r} = p_i^{-1}(w) \cap (\cup_{x \in p^{-1}(v)} [x-r, x+r])$$

在集合 $\underline{M}_i$ 中。子集 $p_i^{-1}(w)_{\leq r}$ 筛选纤维 $p_i^{-1}(w)$。注意到 $p_i^{-1}(v)_{\leq r} = p_i^{-1}(v)$。

Set

$$d_i[r](v,w) = \sum_{x \in p_i^{-1}(v), y \in p_i^{-1}(w), d(x,y) \leq r} d(x,y), \tag{8}$$

并定义

$$d[r](v,w) = \sum_i d_i[r](v,w)$$

对所有 $v, w \in \mathcal{L}$。

数 $d_i[r](v,w)$ 非零，当且仅当存在元素 $x \in p_i^{-1}(v)$ 和 $y \in p_i^{-1}(w)$，使得 $d(x,y) \leq r$，并且 $d[r](v,w) \neq 0$；当且仅当对于某个 i，有$d_i[r](v,w) \neq 0$。

特别地，$d_i[r](v,v)$ 是满足 $d(x,y) \leq r$ 且 $d_i[0](v,v) = 0$ 的 $x, y \in p_i^{-1}(v)$ 之间的距离 $d(x,y)$ 之和。因此，对于 $r > 0$，$d[r](v,v)$ 可能是非平凡的，并且 $d[0](v,v) = 0$。

取纤维 $p_i^{-1}(v)$ 中的所有元素 x，并构造 $\underline{M}_i$ 中的所有区间 $[x-r, x+r]$。其并集

$$N_v[r] = \cup_i (\cup_{x \in p_i^{-1}(v)} p_i[x-r, x+r]) \subset \mathcal{L}. \tag{9}$$

是 $\mathcal{L}$ 中 v 的一个邻域。

注意，$N_v[0] = \{v\}$。此外，对于 $r \leq s$，$N_v[r] \subset N_v[s]$，且 $\cup_r N_v[r] = \mathcal{L}$，因此子集 $N_v[r]$ 筛选单词集合 $\mathcal{L}$。

固定一个正整数 r 后，集合 $N_v[r]$ 是$v \in \mathcal{L}$的一个邻域，而数 $d[r](v,w)$ 是 $w \in N_v[r]$ 的权重。

与第5节一样，UMAP 构造将加权邻域 $(N_v[r], d[r])$、$v \in \mathcal{L}$ 组合起来，形成 UMAP 复形 $V(\mathcal{L}, N[r])$、一个拟度量空间 $(\mathcal{L}, D[r])$，以及射线复形 $R(N[r]) \subset V(\mathcal{L}, D[r])$，它们计算出的簇都相同。

8 抽样

假设通用数据集 $\mathcal{U}$ 具有 ep-度量空间结构，但没有其他信息。

在这种情况下，对于给定点 $x \in \mathcal{U}$，可通过一种基于采样技术并在样本中构造 k-完备邻域的蛮力法来近似（或发现）一个邻域 N_x。

假设 Z 是从 $\mathcal{U}$ 中随机选取的子集（一个样本），并且 Z 是易处理的，即存在一个基数上界 $|Z| \leq M$，使得可用的计算设备能够分析大小至多为 M 的数据集。我们假设 $x \in Z$。

对于这样的子集 Z，可以计算距离函数 $d_x : Z \to [0, \infty)$，其中 $d_x(z) = d(x, z)$，而 d_x 的像 $d_x(Z)$ 定义了区间 $[0, \infty)$ 的一个易处理子集。集合 Z 是距离函数 Z 的纤维的不交并

$$Z = \sqcup_{s \in d_x(Z)} \, d_x^{-1}(s)$$

距离函数 d_x 的

假设 k 是一个满足 $k \leq |Z|$ 的固定正整数。

元素 x 在 Z 中有唯一确定的 k-完备邻域 N，如第 2 节所述；它是满足 $|Z(x, s)| \geq k$ 的最小完备邻域 $Z(x, s)$。邻域 N 是由 $d_x^{-1}(t)$ 构成的纤维的并，对应于 t 的最小值，

对于适当大小的样本集合 $Z_1, \ldots, Z_p$，其中包含 x，可以并行重复此构造，并使用距离函数 $d_x : Z_i \to [0, \infty)$。每个样本 Z_i 都有唯一确定且可计算的 k-完备邻域 N_i，而并集 $\cup_i Z_i$ 中 x 的 k-完备邻域 N，根据引理 10，是较小对象 $\cup_i N_i$中 x 的 k-完备邻域。

调用样本 Z_i有多种方式:

1) 从样本 Z_x中 x 的一个 k-完备邻域 N_x 开始，为每个 $y \in N_x$选择样本 Z_y，并为其关联相应的 k-完备邻域 $N_y \subset Z_y$。并集 $\cup_{y \in N_x} N_y$ 包含一个 k-完备邻域 N'_x，它可能是对宇宙 $\mathcal{U}$中 x 的 k-完备邻域的更好近似。

这一系列步骤是 [2]的 k-最近邻算法的一个类似方法。

2) 对于某个 $V \subset \mathcal{U}$，V 中 x 的 k-完备邻域 N 的确定，可以扩展到 $\mathcal{U}$ 的更大子集，但须满足计算约束；方法是向 V 添加更多易处理的样本。这再次是引理 10 的简单应用。

3) 如果 $Z_1, \ldots, Z_p$ 是在 $\mathcal{U}$ 中一组易处理样本，那么对于任意 $y \in Z$，我们都可以在 $Z = \cup_i Z_i$ 中找到一个 k-完备邻域 N_y。相应的子复形 $V(N_y) \subset V(Z)$ 和 $R(N_y) \subset V(N_y)$ 确定了滤过子复形

$$\cup_{y \in Z} R(N_y) \subset \cup_{y \in Z} V(N_y),$$

由此得到一种 UMAP 风格分析，用于计算 $V(Z)$ 的簇，并近似所有 $V(\mathcal{U})$ 的簇。

这里展示的采样技术完全是蛮力式的。它只近似了点的簇和邻域，并未涵盖整个数据集 $\mathcal{U}$。

在存在全局约束时，可以改进该方法，例如利用 [2] 的局部均匀性假设，生成 k 个最近邻集合，并给出概率估计。

参考资料

[1] David M Blei、Andrew Y Ng 和 Michael I Jordan。Latent dirichlet allocation。机器学习研究期刊，3(Jan):993–1022，2003。[2] Wei Dong、Charikar Moses 和 Kai Li。针对通用相似性度量的高效 k 最近邻图构建。载于第 20 届万维网国际会议论文集，WWW'11，第 577–586 页，New York, 2011。ACM。[3] J.F. Jardine。UMAP 的稳定性。预印本，arXiv: 2011。13430 [math.AT],2020。[4] Leland McInnes。UMAP 文档。github.com/lmcinnes/umap，2023。[5] Leland McInnes、John Healy 和 James Melville。UMAP：用于降维的统一流形近似与投影。预印本，arXiv: 1802。03426 [stat.ML], 2020。[6] Gurjeet Singh、Facundo Memoli 和 Gunnar Carlsson。用于高维数据集分析和三维物体识别的拓扑方法。载于 M. Botsch、R. Pajarola、B. Chen 和 M. Zwicker 编，《Eurographics 基于点的图形学研讨会》。Eurographics 协会，2007。