第 5 章 聚　　类

聚类（clustering）的目的是将一个对象集合按照以下的方式分组：在同一个聚类中的各个对象的相似度比它们与其他聚类中的对象的相似度更高。尽管不同对象的特征不同，如基因的 DNA 序列、社交网络中的成员、自然语言的文本、股票价格的时间序列、计算机断层扫描的医学图像或电子商务平台上的消费产品等，它们各自的聚类任务可能是非常相关的。聚类本身并不是一个特定的算法，而是一个需要解决的任务。聚类可以通过很多算法来实现，但这些算法在对聚类的构成以及如何有效识别聚类的理解上存在显著的差异。本章将介绍著名的基于对象之间的一般**差异性度量**（dissimilarity measure）的 ***k*-均值聚类**（k-means clustering）。首先，该算法会将每个对象分配到与其差异度最小的聚类中心。然后，该算法会通过最小化聚类内部的差异性来重新计算聚类中心。k-均值算法是根据欧几里得设定执行的，其中聚类的中心就是聚类中样本的均值。此外，我们讨论了 k-均值算法基于其他不同的相异性度量的改动，包括 **Levenshtein 距离**（Levenshtein distance）、**Manhattan 范数**（Manhattan norm）、**余弦相似度**（cosine similarity）、**Pearson 相关性**（Pearson correlation）和 **Jaccard 系数**（Jaccard coefficient）。最后，我们将**谱聚类**（spectral clustering）技术应用在了**社区发现**（community detection）问题上。这一应用是建立在研究信息通过社交网络的传播和对相应的转移概率矩阵做谱分析来实现的。

5.1 研究动因：DNA 测序

近年来，人们采集了大量的**基因表达数据**（gene expression data），其中的 DNA 测序问题引起了大量关注。1998 年，冰岛议会决定允许生物制药公司 deCODE Genetics 收集和存储其全国人口的所有健康数据。其法律依据是随后不久批准的生物样本库法案（Act of Biobanks）。在之后的 2003 年，deCODE Genetics 就推出了一个在线版本的 DNA 测序数据库，名为 Íslendingabók，又名“冰岛人之书”（Book of Icelanders）。任何拥有冰岛社会保障号码的人都可以研究他们的家谱，并查看他们与该国其他任何人之间最近的家庭关系。到 2020 年，“冰岛人之书”已经拥有了超过 20 万名注册用户和超过 90 万个 DNA 序列的关联条目，这些用户中包括大多数冰岛人，无论他们是仍然健在还是已经去世。由

于 deCODE 基因表达数据集一直是世界上最大的且支持维护得最好的数据集之一，因此人们侧重使用它做谱系关系的分析。具体地说，人们在基因表达数据中探寻有意义的信息模式和依赖关系，这项任务虽然困难，但却是必不可少的，因为它可以为假设检验提供基础。谱系分析的一种解决方案是将 DNA 序列相互关联，而非每次都独立地重新访问每个新的 DNA 序列。识别具有相似 DNA 模式的同质群体的任务通常是由**聚类**（clustering）实现的。DNA 序列的聚类已被证明可以很好地应用于揭示基因表达数据中内在的自然结构，其中包括基因功能、细胞过程和细胞类型。另一个好处是它能够更好地帮助人们理解基因同源性，这一点在设计疫苗的过程中非常重要。DNA 序列的差异性度量是对基因表达数据聚类的关键所在。简单来说，DNA 链可以表示为一个字符串，即一个由以下字母表所构成的有限字符序列：

$$\Sigma = \{\text{Adenin(A), Guanin(G), Thymin(T), Cytosin(C)}\}。^{①}$$

用 Σ^s 表示长度为 s 的字符串集合，所有有限字符串（无论长度）的集合用克林星（Kleene star）算子表示：

$$\Sigma^* = \bigcup_{s\in\mathbb{N}\cup\{0\}} \Sigma^s。$$

对于任意两个 DNA 序列 $x, y \in \Sigma^*$，我们需要定义相应的差异性度量 $d(x, y)$。为此，一个合理的选项是 Levenshtein 距离，它可以衡量两个序列之间的编辑距离：

$$d(x, y) = \text{lev}(x, y)。$$

Levenshtein 距离（Levenshtein distance）$\text{lev}(x, y)$ 表示将字符串 x 更改为字符串 y 所需的最小单字符编辑（插入、删除或替换）次数。它以 Levenshtein 的名字命名。为了更好地理解这个定义，让我们计算下面两个字符串之间的 Levenshtein 距离

$$x = \text{ACCGAT}, \quad y = \text{AGCAT}。$$

从 x 编辑获得 y 的一种可能的操作是将第 2 个字符 C 替换为 G，将第 4 个字符 G 替换为 A，将第 5 个字符 A 替换为 T，并删除最后一个字符 T：

A	C	C	G	A	T
	↑		↑	↑	↑
	G		A	T	×

总共需要进行 4 次编辑，但其实我们可以做得更好。实际上，我们可以将第 2 个字符 C 替换为 G，并删除第 4 个字符 G：

① 译者注：这里的 A、T、G、C 代表的分别是四种碱基，即腺嘌呤（Adenin）、鸟嘌呤（Guanin）、胸腺嘧啶（Thymin）和胞嘧啶（Cytosin）。

$$\begin{array}{cccccc} \mathrm{A} & \mathrm{C} & \mathrm{C} & \mathrm{G} & \mathrm{A} & \mathrm{T} \\ & \uparrow & & \uparrow & & \\ & \mathrm{G} & & \times & & \end{array}$$

因此，如果使用 Levenshtein 距离，有 $\mathrm{lev}(x, y) = 2$。现在，我们将注意力转向更一般的问题：如何通过给定的差异性度量来有效地将对象聚类？

5.2　研究结果

5.2.1　*k*-均值算法

我们首先描述和分析用于聚类的基本 k-均值算法。

1. 总差异性

假定已知来自一个数据集中的 n 个对象 $x_1, \cdots, x_n$。它们需要被分为 k 个同质的聚类 $C_1, \cdots, C_k$，它们是互不相交的，且其并集可以构成一个划分，即

$$\bigcup_{l=1}^{k} C_l = \{1, \cdots, n\}。$$

聚类的同质性意味着它的成员相互之间较为近似。同时，不同聚类间的成员又应该具有显著的可区分性。为了落实这个想法，假设我们可以有效地测量任意两个对象 x 和 y 之间的**差异度**（dissimilarity）$d(x, y)$。这个度量是执行聚类的基础。为此，我们将聚类 $C_1, \cdots, C_k$ 与它们的聚类中心 $z_1, \cdots, z_k$ 关联起来。聚类中心 z_l 可以代表聚类 C_l，因此聚类内的差异度可以计算为

$$d(C_l, z_l) = \sum_{i \in C_l} d(x_i, z_l)。$$

所以，总的差异度可以由下式给出

$$d(C, z) = \sum_{l=1}^{k} d(C_l, z_l),$$

其中聚类和其相应的聚类中心可以表示为

$$C = (C_1, \cdots, C_k), \quad z = (z_1, \cdots, z_k)。$$

我们的目标是找到总差异度最小的聚类划分 C 和这个划分的相应聚类中心 z：

$$\min_{C, z} d(C, z)。 \tag{5.1}$$

2. 朴素 k-均值方法

由于优化问题式 (4.8) 的组合性质，所以它很难解决。然而，目前有一个简单的迭代方案，它可以交替地更新聚类划分和聚类中心。结果证明，这种方法非常有效。这个算法由贝尔实验室的 Lloyd 提出 [Lloyd, 1982]，起初是一种用来做脉冲编码调制的技术，而这个想法最早可以追溯到 [Steinhaus et al., 1956]。我们下面正式介绍 ***k*-均值聚类算法**（k-means clustering）：

（1）下一个迭代中的每个聚类均由与前一个迭代的聚类中心差异度最低的对象组成，即对于所有 $l = 1, \cdots, k$，有：

$$C_l(t+1) = \{i \in \{1, \cdots, n\} | d(x_i, z_l(t)) \leqslant d(x_i, z_{l'}(t)) \text{ 对任意} l' = 1, \cdots, k\}。$$

（2）下一个迭代的**聚类中心**（centers）的选择会将新形成的聚类内的差异度最小化，即对于所有 $l = 1, \cdots, k$，有

$$z_l(t+1) \in \arg\min_z d(C_l(t+1), z)。$$

k-均值聚类由一系列从第 t 次到第 $t+1$ 次的迭代构成，每次迭代会进行聚类-中心的交替更新。

$$(C(t), z(t)) \overset{(1)}{\longmapsto} (C(t+1), z(t)) \overset{(2)}{\longmapsto} (C(t+1), z(t+1))。$$

现在证明总差异度没有增加：

$$\begin{aligned} d(C(t+1), z(t+1)) &= \sum_{l=1}^{k} d(C_l(t+1), z_l(t+1)) \overset{(2)}{\leqslant} \sum_{l=1}^{k} d(C_l(t+1), z_l(t)) \\ &= \sum_{l=1}^{k} \sum_{i \in C_l(t+1)} d(x_i, z_l(t)) \overset{(1)}{\leqslant} \sum_{l'=1}^{k} \sum_{i \in C_{l'}(t)} d(x_i, z_{l'}(t)) \\ &= \sum_{l'=1}^{k} d(C_{l'}(t), z_{l'}(t)) = d(C(t), z(t))。 \end{aligned}$$

由于划分的数量是有限的，因此总差异度的序列 $(d(C(t), z(t)))_{t \in \mathbb{N}}$ 也是一个有限数量的非递增序列。因此，k-均值方法将在有限次数的迭代步骤后停止，收敛至优化问题式(5.1)的**局部**（local）最小值。k-均值方法向**全局**（global）最小值的收敛的关键在于聚类中心的初始化，并且在通常情况下该方法并不能保证会得到全局最小值。我们将通过图 5.1的示例看到，除非聚类中心的初始化非常恰当，否则 k-均值算法将很难达成收敛至全局最小化，至少在欧几里得设定中将是如此。为了克服这个障碍，一个简单的方法是可以重复多次执行 k-均值法，每次将聚类中心初始化为不同的值。Arthur 和 Vassilvitskii[Vassilvitskii and

Arthur, 2007] 提出了另一种称为 **k-均值 ++**（k-means++）的初始化策略。在这种方法中，聚类中心是随机选择的，其概率与已经被选中的中心的距离平方成正比。

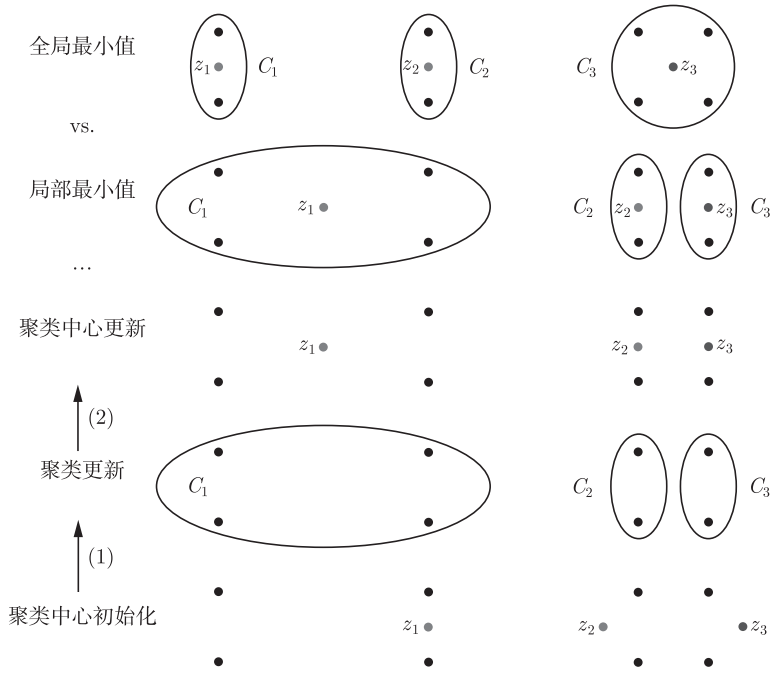

图 5.1　k-均值聚类

3. 欧几里得设定

我们考虑欧几里得设定下的 k-均值聚类。为此，假设数据点 $x_1, \cdots, x_n \in \mathbb{R}^m$ 描述了对象的 m 个特征，对于任意两个对象 $x, y \in \mathbb{R}^m$，它们之间的差异度度量由**欧几里得距离**（Euclidean distance）给出：

$$d(x, y) = \|x - y\|_2^2 。$$

分别研究由（1）和（2）给出的聚类及其中心的结构。为简单起见，我们省略了迭代次数 $t+1$。首先，我们来具体表述（2）中的中心的更新。为此，需要找到可以最小化集群 C_l 内部的差异度的中心 z_l：

$$\min_z \sum_{i \in C_l} \|x_i - z\|_2^2 。$$

最优的必要条件为：

$$\nabla_z \left(\sum_{i \in C_l} \|x_i - z\|_2^2 \right) = 0,$$

或者与其等价的：

$$2 \cdot \sum_{i \in C_l} (z - x_i) = 0。$$

此式也就是众所周知的**样本均值**（sample mean）公式：

$$z_l = \frac{1}{|C_l|} \sum_{i \in C_l} x_i,$$

其中，$|C_l|$ 表示 C_l 内成员的个数。我们可以得出结论：（2）式中的聚类中心应该选取聚类内数据点的样本均值。而其实这就是 k-均值这个名称的来源。此外，请注意，可以通过聚类中心得到一个 **Voronoi 图**（Voronoi diagram），这个图将 $\mathbb{R}^m$ 分解为 k 个凸单元，每个凸单元都包含最近中心为 z_l 的空间区域，其中 $l = 1, \cdots, k$。由于由（1）得到的聚类具有以下形式：

$$C_l = \{i \in \{i, \cdots, n\} | \, \|x_i - z_l\|_2 \leqslant \|x_i - z_{l'}\|_2 \text{ 对全部} l' = 1, \cdots, k\},$$

那么数据点 x_i，$i \in C_l$ 也会落在相应的 Voronoi 单元之内，如图 5.2 所示。

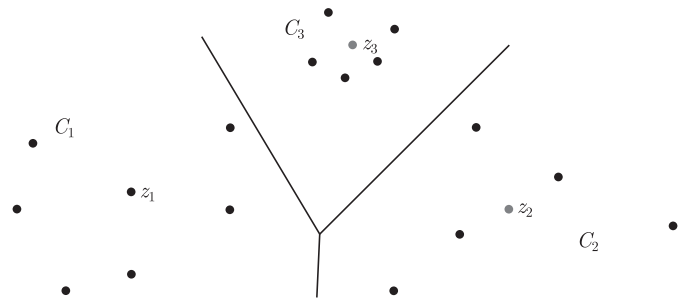

图 5.2　Voronoi 图

5.2.2　谱聚类

下面考虑一个 k-均值算法的应用，即用于社区发现的谱聚类，参见 [Mathar et al., 2020]。

1. 社区发现

社区发现（community detection）是理解复杂社交网络结构并最终从中提取有用信息的关键。社区发现背后的动机是多种多样的。它可以帮助品牌和商家了解人们对其产品的不同意见。品牌方可以针对兴趣相似且地理位置相近的特定人群量身定制解决方案，进而提高服务的绩效。识别购买关系网络中的相似客户，可以帮助在线零售商建立有效的推荐机制，从而更好地引导客户去浏览商品推荐列表，增加零售商的商机。下面对社区发现问

题进行数学建模。为此，假设**社交网络**（social network）中有 n 个人可以相互交流通信，如图 5.3 所示。假设对于社交网络中的任意两个人 i 和 j，它们的**连接值**（linkage）$w_{ij} \geqslant 0$ 均为已知。w_{ij} 表示 i 和 j 两人之间的连接强度的权重。这个权重可能为每天通话的平均时间等，因此对于所有的 $i, j = 1, \cdots, n$，有以下关系成立：

$$w_{ij} = w_{ji}。$$

简洁起见，我们定义在非负元素上的 $n \times n$ 维对称连接矩阵：

$$\boldsymbol{W} = (w_{ij})。$$

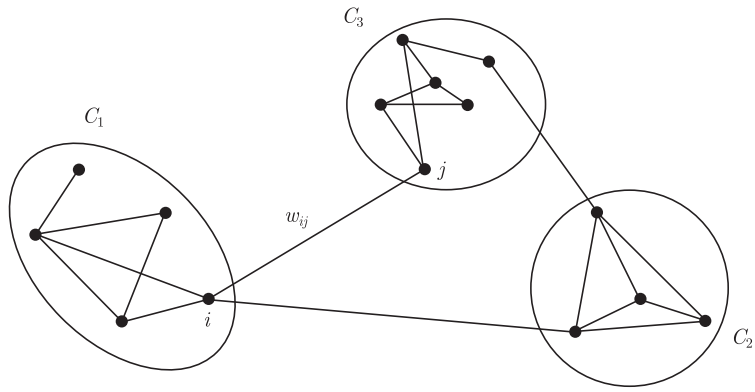

图 5.3　社区发现

为了将所有的人聚类到 k 个组中，需要一个适当的差异性度量。通常，直接使用连接值的信息是不够的。事实上，尽管有时人们彼此交谈很多，但最终还是意见相左，因为他们各自也会与其他人有联系。而谱聚类的主要思想则是观察分析社会网络中的**信息扩散**（diffusion of information）。粗略来说，如果两个人一段时间后在社交网络上产生了大致相同的信息足迹，那么我们就认为他们是相似的。换句话说，人的差异度是通过他们以不同方式影响和操纵他人的意见的能力来衡量的。下面用数学的术语来表达这个想法。

2. 信息扩散

首先，信息从第 j 个人扩散到第 i 个人的概率为：

$$p_{ij} = \frac{w_{ij}}{\sum_{i=1}^{n} w_{ij}}。$$

对于 $n \times n$ 维度的矩阵 $\boldsymbol{P} = (p_{ij})$，有：

$$\boldsymbol{P} = \boldsymbol{W} \cdot \boldsymbol{D}^{-1},$$

其中关于度的对角矩阵由下式给出

$$\boldsymbol{D}=\operatorname{diag}\left(\sum_{i=1}^{n} w_{ij}, j=1,\cdots,n\right)。$$

注意，$\boldsymbol{P}$ 是构造出的**随机矩阵**（stochastic matrix），即它的元素是非负的，并且它的列的总和为 1：

$$\boldsymbol{P}\geqslant 0,\quad \boldsymbol{e}^{\top}\cdot\boldsymbol{P}=\boldsymbol{e}^{\top},$$

其中 $\boldsymbol{e}=(1,\cdots,1)^{\top}$ 表示 n 维的全 1 向量。现在，我们描述信息扩散的过程。为此，假设八卦消息是从第 j 个人开始的：

$$\boldsymbol{x}_j(0)=\boldsymbol{e}_j,$$

其中 $\boldsymbol{e}_j$ 是 $\mathbb{R}^n$ 的第 j 个坐标向量，即 $\boldsymbol{e}_j$ 的第 j 个元素等于 1，而其余元素为零。这意味着社交网络中的所有人，除了第 j 个人，都还暂时不知道这条新闻。一旦这个信息被传递给第 j 个人的朋友后，它就会被一传十十传百地传播到整个社交网络中。因此，这种信息的扩散是由以下的更新方程给出的：

$$\boldsymbol{x}_j(t)=\boldsymbol{P}\cdot\boldsymbol{x}_j(t-1),$$

其中 $\boldsymbol{x}_j(t-1)\in\Delta$ 可以被看作社交网络上的第 j 个人在时间 $t-1$ 时刻引发的信息足迹。回想我们在这里使用过的单纯形：

$$\Delta=\{\boldsymbol{x}\in\mathbb{R}^n|\boldsymbol{x}\geqslant 0\ 且\boldsymbol{e}^{\top}\cdot\boldsymbol{x}=1\}。$$

这里，我们提出的动态过程是良定的，因为在零时刻，第 j 个人从信息足迹是从 $\boldsymbol{x}_j(0)\in\Delta$ 开始的。然后，$\boldsymbol{x}(t)$ 也可以构成一个社交网络上的信息足迹，即 $\boldsymbol{x}_j(t)\in\Delta$。这个结论可以通过归纳法来证明。由于矩阵 $\boldsymbol{P}$ 是随机矩阵，我们有：

$$\boldsymbol{x}_j(t)=\underbrace{\boldsymbol{P}}_{\geqslant 0}\cdot\underbrace{\boldsymbol{x}_j(t-1)}_{\geqslant 0}\geqslant 0,$$

以及

$$\boldsymbol{e}^{\top}\cdot\boldsymbol{x}_j(t)=\underbrace{\boldsymbol{e}^{\top}\cdot\boldsymbol{P}}_{=\boldsymbol{e}^{\top}}\cdot\boldsymbol{x}_j(t-1)=\boldsymbol{e}^{\top}\cdot\boldsymbol{x}_j(t-1)=1。$$

具体地，信息的扩散可以写为

$$\boldsymbol{x}_j(t)=\boldsymbol{P}\cdot\underbrace{\boldsymbol{x}_j(t-1)}_{=\boldsymbol{P}\cdot\boldsymbol{x}_j(t-2)}=\boldsymbol{P}^2\cdot\boldsymbol{x}_j(t-2)=\cdots=\boldsymbol{P}^t\cdot\boldsymbol{x}_j(0)=\boldsymbol{P}^t\cdot\boldsymbol{e}_j。$$

为了计算扩散矩阵 $\boldsymbol{P}$ 的 t 次方，我们将先推导出它的谱分解。

3. 谱分解

我们从辅助矩阵的对角化开始：

$$\boldsymbol{S}=\boldsymbol{D}^{-1/2}\cdot\boldsymbol{W}\cdot\boldsymbol{D}^{-1/2},$$

其中

$$\boldsymbol{D}^{-1/2}=\operatorname{diag}\left(\frac{1}{\sqrt{\sum_{i=1}^{n}w_{ij}}},j=1,\cdots,n\right)。$$

注意，矩阵 $\boldsymbol{S}$ 是对称的，因为对于它的转置，有：

$$\boldsymbol{S}^{\top}=\left(\boldsymbol{D}^{-1/2}\cdot\boldsymbol{W}\cdot\boldsymbol{D}^{-1/2}\right)^{\top}=\left(\boldsymbol{D}^{-1/2}\right)^{\top}\cdot\boldsymbol{W}^{\top}\cdot\left(\boldsymbol{D}^{-1/2}\right)^{\top}=\boldsymbol{D}^{-1/2}\cdot\boldsymbol{W}\cdot\boldsymbol{D}^{-1/2}=\boldsymbol{S}。$$

考虑对称矩阵 $\boldsymbol{S}$ 的特征向量 $\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n\in\mathbb{R}^n$，以及它们对应的特征值 $\lambda_1,\cdots,\lambda_n\in\mathbb{R}$。根据定义，对所有 $r=1,\cdots,n$，有：

$$\boldsymbol{S}\cdot\boldsymbol{v}_r=\lambda_r\cdot\boldsymbol{v}_r。$$

等价地，有以下矩阵形式：

$$\boldsymbol{S}\cdot\boldsymbol{V}=\boldsymbol{V}\cdot\boldsymbol{\Lambda},$$

其中矩阵 $\boldsymbol{V}=(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n)$ 的列由 $\boldsymbol{S}$ 的特征向量构成，而对角矩阵 $\boldsymbol{\Lambda}=\operatorname{diag}(\lambda_1,\cdots,\lambda_n)$ 的对角线上元素为 $\boldsymbol{S}$ 的特征值。为简单起见，假设 $\boldsymbol{S}$ 的所有特征值彼此不同，并且按降序排列：

$$|\lambda_1|>\cdots>|\lambda_n|。$$

于是，我们有

$$\lambda_s\cdot\boldsymbol{v}_r^{\top}\cdot\boldsymbol{v}_s=\boldsymbol{v}_r^{\top}\cdot\underbrace{\lambda_s\cdot\boldsymbol{v}_s}_{=\boldsymbol{S}\cdot\boldsymbol{v}_s}=\boldsymbol{v}_r^{\top}\cdot\boldsymbol{S}\cdot\boldsymbol{v}_s=\left(\boldsymbol{S}^{\top}\cdot\boldsymbol{v}_r\right)^{\top}\cdot\boldsymbol{v}_s=\underbrace{(\boldsymbol{S}\cdot\boldsymbol{v}_r)^{\top}}_{=\lambda_r\cdot\boldsymbol{v}_r^{\top}}\cdot\boldsymbol{v}_s=\lambda_r\cdot\boldsymbol{v}_r^{\top}\cdot\boldsymbol{v}_s。$$

由于我们假设 $\lambda_r\neq\lambda_s$，所以我们从这里推导出对于所有 $r\neq s$，$\boldsymbol{v}_r^{\top}\cdot\boldsymbol{v}_s=0$。一般来说，我们还可以进一步假设这些特征向量使用了欧几里得范数的归一化，即对于所有 $r=1,\cdots,n$，$\|\boldsymbol{v}_r\|_2=1$。总之，$\boldsymbol{S}$ 的特征向量是两两正交的：

$$\boldsymbol{v}_r^{\top}\cdot\boldsymbol{v}_s=\begin{cases}1, & r=s\\ 0, & r\neq s。\end{cases}$$

与此等价，$\boldsymbol{V}$ 是一个**正交矩阵**（orthogonal matrix）：

$$\boldsymbol{V}^{\top}\cdot\boldsymbol{V}=\boldsymbol{V}\cdot\boldsymbol{V}^{\top}=\boldsymbol{I},$$

其中 $\boldsymbol{I}$ 表示单位矩阵。综上，将 $\boldsymbol{S}$ 对角化：

$$\boldsymbol{S}=\boldsymbol{S}\cdot\underbrace{\boldsymbol{V}\cdot\boldsymbol{V}^{\top}}_{=\boldsymbol{I}}=\underbrace{\boldsymbol{S}\cdot\boldsymbol{V}}_{=\boldsymbol{V}\cdot\boldsymbol{\Lambda}}\cdot\boldsymbol{V}^{\top}=\boldsymbol{V}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{V}^{\top}。$$

现在，就可以谱形式表示扩散矩阵 $\boldsymbol{P}$ 了：

$$\boldsymbol{P}=\boldsymbol{W}\cdot\boldsymbol{D}^{-1}=\underbrace{\boldsymbol{D}^{1/2}\cdot\boldsymbol{D}^{-1/2}}_{=\boldsymbol{I}}\cdot\boldsymbol{W}\cdot\underbrace{\boldsymbol{D}^{-1/2}\cdot\boldsymbol{D}^{-1/2}}_{=\boldsymbol{D}^{-1}}=\boldsymbol{D}^{1/2}\cdot\underbrace{\boldsymbol{D}^{-1/2}\cdot\boldsymbol{W}\cdot\boldsymbol{D}^{-1/2}}_{=\boldsymbol{S}}\cdot\boldsymbol{D}^{-1/2}$$

$$=\boldsymbol{D}^{1/2}\cdot\underbrace{\boldsymbol{S}}_{=\boldsymbol{V}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{V}^{\top}}\cdot\boldsymbol{D}^{-1/2}=\underbrace{\boldsymbol{D}^{1/2}\cdot\boldsymbol{V}}_{=\boldsymbol{\Phi}}\cdot\boldsymbol{\Lambda}\cdot\underbrace{\boldsymbol{V}^{\top}\cdot\boldsymbol{D}^{-1/2}}_{=\boldsymbol{\Psi}^{\top}}=\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{\Psi}^{\top}。$$

矩阵 $\boldsymbol{\Psi}=(\boldsymbol{\psi}_1,\cdots,\boldsymbol{\psi}_n)$ 和 $\boldsymbol{\Phi}=(\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_n)$ 是**双正交的**（bi-orthonormal），即它们的列彼此正交：

$$\boldsymbol{\Psi}^{\top}\cdot\boldsymbol{\Phi}=\underbrace{\boldsymbol{V}^{\top}\cdot\boldsymbol{D}^{-1/2}}_{=\boldsymbol{\Psi}^{\top}}\cdot\underbrace{\boldsymbol{D}^{1/2}\cdot\boldsymbol{V}}_{=\boldsymbol{\Phi}}=\boldsymbol{V}^{\top}\cdot\underbrace{\boldsymbol{D}^{-1/2}\cdot\boldsymbol{D}^{1/2}}_{=\boldsymbol{I}}\cdot\boldsymbol{V}=\boldsymbol{V}^{\top}\cdot\boldsymbol{V}=\boldsymbol{I},$$

或者，等价地：

$$\boldsymbol{\psi}_r^{\top}\cdot\boldsymbol{\phi}_s=\begin{cases}1, & r=s,\\ 0, & r\neq s。\end{cases}$$

4. 扩散图

通过使用矩阵 $\boldsymbol{P}$ 的谱分解，我们得到了它的 t 次幂的简单表示：

$$\boldsymbol{P}^t=\underbrace{\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{\Psi}^{\top}}_{=\boldsymbol{P}}\cdot\underbrace{\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{\Psi}^{\top}}_{=\boldsymbol{P}}\cdot\ \cdots\ \cdot\underbrace{\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{\Psi}^{\top}}_{=\boldsymbol{P}}\cdot\underbrace{\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{\Psi}^{\top}}_{=\boldsymbol{P}}$$

$$=\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}\cdot\underbrace{\boldsymbol{\Psi}^{\top}\cdot\boldsymbol{\Phi}}_{=\boldsymbol{I}}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{\Psi}^{\top}\cdot\ \cdots\ \cdot\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}\cdot\underbrace{\boldsymbol{\Psi}^{\top}\cdot\boldsymbol{\Phi}}_{=\boldsymbol{I}}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{\Psi}^{\top}=\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}^t\cdot\boldsymbol{\Psi}^{\top}。$$

继续探讨由第 j 个人引起的信息扩散：

$$\boldsymbol{x}_j(t)=\boldsymbol{P}^t\cdot\boldsymbol{e}_j=\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}^t\cdot\boldsymbol{\Psi}^{\top}\cdot\boldsymbol{e}_j=\boldsymbol{\Phi}\cdot\boldsymbol{\Lambda}^t\cdot\begin{pmatrix}(\boldsymbol{\psi}_1)_j\\ \vdots\\ (\boldsymbol{\psi}_n)_j\end{pmatrix}=\boldsymbol{\Phi}\cdot\begin{pmatrix}\lambda_1^t\cdot(\boldsymbol{\psi}_1)_j\\ \vdots\\ \lambda_n^t\cdot(\boldsymbol{\psi}_n)_j\end{pmatrix}。$$

下面将 t 时刻的**扩散图**（diffusion map）定义为

$$\boldsymbol{F}_j(t)=\begin{pmatrix}\lambda_1^t\cdot(\boldsymbol{\psi}_1)_j\\ \vdots\\ \lambda_n^t\cdot(\boldsymbol{\psi}_n)_j\end{pmatrix},$$