模糊等价关系与核函数之间的关系及其在动态模糊聚类中的应用*

统计 91 董晟渤　　统计 91 霍梦钰　　统计 91 赵梓萌

2193510853　　2191421886　　2196123690

西安交通大学数学与统计学院

日期：2021 年 12 月

摘　　要

核函数, 是可以通过非线性变换, 转换为 Hilbert 空间中内积的二元函数. 基于此特性, 其具有将学习、优化及分类等线性模型转换为非线性的应用能力, 且已在机器学习、数据挖掘及计算机视觉化呈现中得到了较为成功的实现.

等价关系, 是模糊数学中的重要概念, 其为界限不明晰的数据的聚类划分提供了坚实的理论依据. 在本文中, 我们基于 t 模的概念定义了 T-等价关系, 给出了几种 t 模的形式并讨论了由其定义的 T-等价关系的强弱.

在具体的聚类实践中, 一个重要的问题是核函数的选取及设计, 根据现有研究, 核函数与模糊等价关系之间具有紧密的连结, 本文对此关系进行了总结及证明. 首先, 指出了所有可以作为等价关系的核函数都是 $T_{\cos}$-等价的, 进一步地, 任何核函数均可用 t 模的双蕴含关系表示. 另一方面, 针对核函数的设计问题, 利用等价关系可以实现. 在研究了几类 t 模性质的基础上, 我们证明了 T_M-等价关系均为核函数. 此外, 我们研究了性质并不那么理想的 t 模 T_L、T_P, 指明了其生成核函数的方法, 并提出用这两个 t 模生成 T-等价关系的方法, 形成了完整的理论体系.

最后, 我们给出了使用基于核函数与 T-等价关系的动态模糊聚类实例, 针对不同的核函数及 T-等价关系, 绘制出了聚类结果图、动态效果图、轮廓数值图及热力图并对结果做了简单分析, 证实了此部分理论的极高应用性.

关键词: 核函数; t 模; T-等价关系; 动态模糊聚类

*2021-2022 学年第一学期, 课程名称: 不确定数据分析, 指导老师: 张红英.

目录

1 背景知识: 核函数、t 模和 T-等价关系

1.1 核函数及其封闭性

核函数是一个二元函数, 是由 Hilbert 空间上的内积的概念所推广得到的.

定义 1.1 (核函数)**.** 设 $\mathscr{X}$ 是非空集, 函数 $k:\mathscr{X}\times\mathscr{X}\to\mathbb{R}$ 满足

(1) 对称性, 也即对任意的 $x,y\in\mathscr{X}$, 有 $k(x,y)=k(y,x)$;

(2) 半正定性, 也即对任意的 $n\in\mathbb{N}$, $x_1,x_2,\cdots,x_n\in\mathscr{X}$, $c_1,c_2,\cdots,c_n\in\mathbb{R}$, 有

$$\sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_jk(x_i,x_j)\geq 0,$$

则称其为 $\mathscr{X}$ 上的**核函数**.

例 1.2 (欧氏空间上的核函数)**.** 设 $\mathscr{X}=\mathbb{R}^n$, $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$, 则

$$k_1(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}^T\boldsymbol{y},\quad k_2(\boldsymbol{x},\boldsymbol{y})=\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})^T\right\}$$

是核函数, 分别称为**线性核函数**和 **Gauss 核函数**, 其中 $\boldsymbol{\Sigma}$ 是协方差矩阵. 特别地, 当 $\boldsymbol{\Sigma}$ 是对角线上都为 σ^2 的对角矩阵时, Gauss 核函数可以写成

$$k_2(\boldsymbol{x},\boldsymbol{y})=\exp\left\{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right\}.$$

在 $\mathbb{R}^2$ 上, 作出例1.2中的核函数的图像如图1所示. 容易发现, Gauss 核函数从 $y=x$ 的方向看过去, 就像是 Gauss 分布的密度函数.

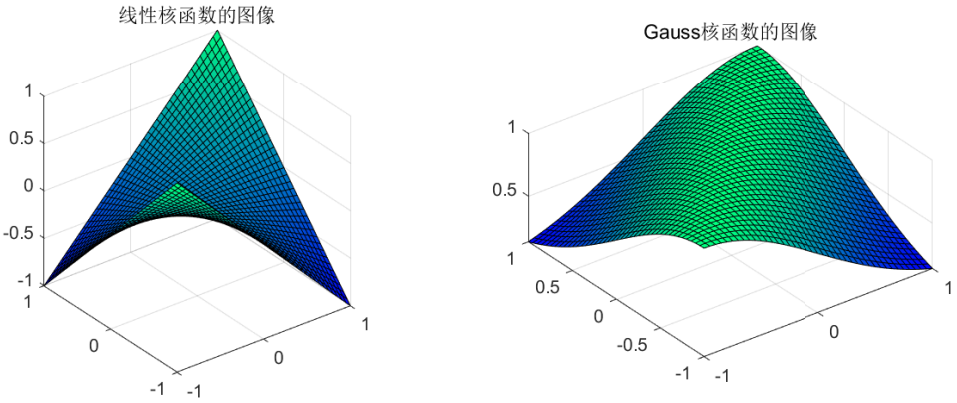


图 1: $\mathbb{R}^2$ 上的线性核函数和 Gauss 核函数的图像, $\sigma=1$

除了上面所提到的核函数以外, $\mathbb{R}^n$ 上还有许多核函数, 见表1. 容易发现, 这些核函数都满足 $k(\boldsymbol{x},\boldsymbol{x})=1(\forall\boldsymbol{x}\in\mathbb{R}^n)$. 我们后面将会指出, 满足这一条件的核函数是非常有用的. 为了方便理解, 我们作出这几个核函数的图像如图2所示.

表 1: 更多核函数的例子

名称	表达式
Laplace 核函数	$k(\boldsymbol{x},\boldsymbol{y}) = \exp\left\{-\dfrac{\|\boldsymbol{x}-\boldsymbol{y}\|}{\sigma}\right\}$
二次有理核函数	$k(\boldsymbol{x},\boldsymbol{y}) = 1 - \dfrac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^2 + c}$
多元二次核函数	$k(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\|\boldsymbol{x}-\boldsymbol{y}\|^2 + c^2}$
逆多元二次核函数	$k(\boldsymbol{x},\boldsymbol{y}) = \dfrac{1}{\sqrt{\|\boldsymbol{x}-\boldsymbol{y}\|^2 + c^2}}$

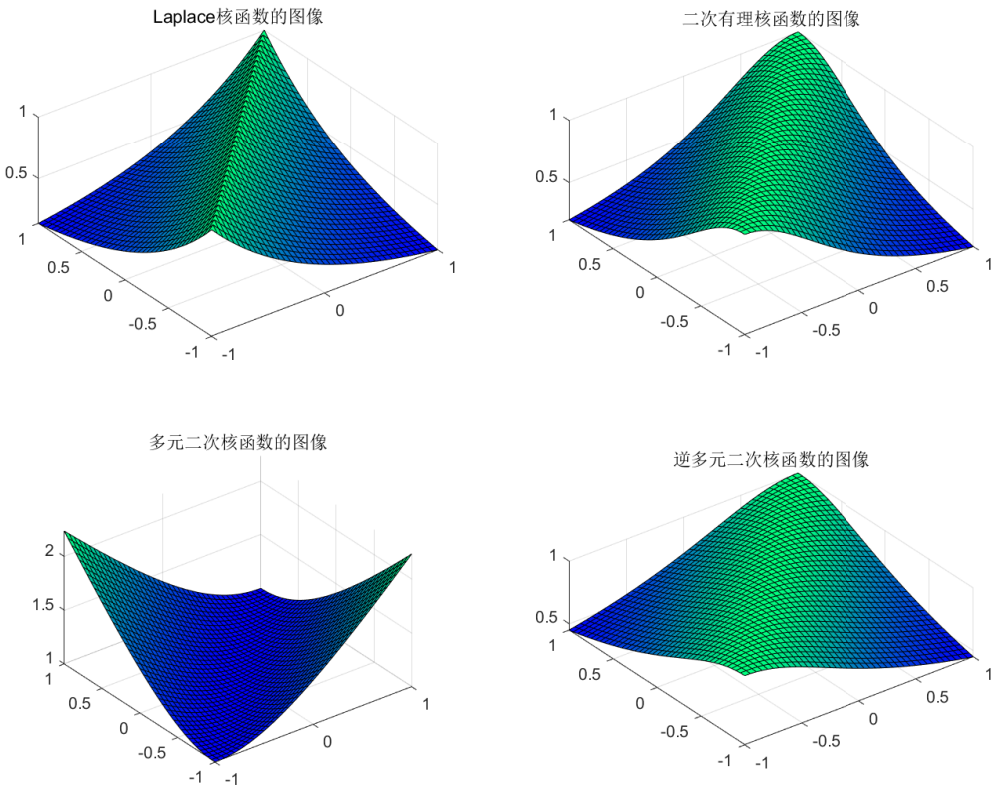


图 2: $\mathbb{R}^2$ 上表1中的核函数的图像, $\sigma = 1, c = 1$

从已知的核函数可以生成新的核函数, 但是所进行的操作必须保持半正定性. 定理1.3给出了这样的操作的特征.

定理 1.3 (核的封闭性, C.H.FitzGerald, 1995)**.** 设 $f:\mathbb{R}^n\to\mathbb{R}$, 定义 $k:\mathscr{X}\times\mathscr{X}\to\mathbb{R}$,

$$k(x,y)=f(k_1(x,y),k_2(x,y),\cdots,k_n(x,y)),$$

其中 $k_1,k_2,\cdots,k_n$ 是 $\mathscr{X}\times\mathscr{X}$ 上的核函数, 则 k 是核函数当且仅当 f 是 $\mathbb{C}^n$ 上形如

$$f(x_1,x_2,\cdots,x_n)=\sum_{r_1\geq 0}\sum_{r_2\geq 0}\cdots\sum_{r_n\geq 0}c_{r_1,r_2,\cdots,r_n}x_1^{r_1}x_2^{r_2}\cdots x_n^{r_n}$$

的整函数在 $\mathbb{R}^n$ 上的限制, 其中对所有的非负指数 $r_1,r_2,\cdots,r_n$, 有 $c_{r_1,r_2,\cdots,r_n}\geq 0$.

证明. 见参考文献 [3]. □

1.2 t 模、蕴含与双蕴含

t 模又称三角模, 可以认为是乘法运算的推广, 见定义1.4.

定义 1.4 (t 模与 Archimedes 模)**.** 设函数 $T:[0,1]^2\to[0,1]$ 满足

(1) 交换律, 也即对任意的 $x,y\in[0,1]$, 有 $T(x,y)=T(y,x)$;

(2) 结合律, 也即对任意的 $x,y,z\in[0,1]$, 有 $T(x,T(y,z))=T(T(x,y),z)$;

(3) 单调性, 也即对任意的 $x,y,z\in[0,1]$, 设 $y\leq z$, 则 $T(x,y)\leq T(x,z)$;

(4) 边界条件, 也即对任意的 $x\in[0,1]$, 有 $T(x,1)=T(1,x)=x$,

则称 T 是 t **模**或**三角模**. 如果 T 是连续的, 且对任意的 $x\in(0,1)$, 都有

$$T(x,x)<x,$$

则称 T 是 **Archimedes 模**.

根据结合律, 可以按照递推式

$$T_n(x_1,x_2,\cdots,x_n)=T(x_1,T_{n-1}(x_2,\cdots,x_n))$$

得到 n 元的 t 模 $T:[0,1]^n\to[0,1]$. 首先, 我们注意到取最小值的操作就是一个 t 模, 这便得到了最简单的 t 模的例子, 见例1.5.

例 1.5 (T_M)**.** 定义 $T_M(x,y)=\min\{x,y\}$, 则 T_M 满足交换律、结合律、单调性和边界条件, 从而 T_M 是 t 模; 但是, T_M 不是 Archimedes 模.

例1.5给出了一个 t 模的例子, 并且发现其不是 Archimedes 模. 什么样的 t 模才是 Archimedes 模呢? 定理1.6给出了 Archimedes 模的表示.

定理 1.6 (Ling, 1965)**.** *设 $T:[0,1]^2\to[0,1]$ 是 t 模, 则 T 是 Archimedes 模当且仅当存在连续的严格递减函数 $f:[0,1]\to[0,\infty]$, 满足 $f(1)=0$, 且对任意的 $x,y\in[0,1]$,*

$$T(x,y)=f^{-1}\left(\min\{f(x)+f(y),f(0)\}\right).$$

证明. 见参考文献 [4]. □

例 1.7 (T_L、$T_{\cos}$ 和 T_P)**.** 在定理1.6中, 首先令 $f(x)=1-x$, 则可得到 **Lukasiewcz** t **模**

$$T_L(x,y)=\max\{x+y-1,0\}.$$

再令 $f(x)=\arccos x$, 记 $x=\cos\alpha, y=\cos\beta$, 则可得到

$$\begin{aligned}T_{\cos}(x,y)&=\cos\left(\min\left\{\alpha+\beta,\frac{\pi}{2}\right\}\right)\\&=\max\left\{\cos(\alpha+\beta),0\right\}\\&=\max\left\{xy-\sqrt{1-x^2}\cdot\sqrt{1-y^2},0\right\}.\end{aligned}$$

接下来, 记 $g(x)=\exp\{-f(x)\}$, 则 $g:[0,1]\to[0,1]$ 是严格递减函数, $g(1)=1$, 且有

$$T(x,y)=g^{-1}\left(\max\{g(x)g(y),g(0)\}\right).$$

在此基础上, 令 $g(x)=x$, 可以得到**乘积** t **模**

$$T_P(x,y)=xy;$$

根据定理1.6, Archimedes 模和 T_P 或 T_L 之一是同构的, 前者称为**非严格的**, 后者称为**严格的**. 特别地, $T_{\cos}$ 是非严格的.

命题 1.8. *对任意的 $x,y\in[0,1]$, 有*

$$T_{\cos}(x,y)\le T_L(x,y)\le T_P(x,y)\le T_M(x,y).$$

证明. 对任意的 $x,y\in[0,1]$, 首先,

$$\begin{aligned}&xy-\sqrt{1-x^2}\cdot\sqrt{1-y^2}\le x+y-1\\\iff\ &1-x-y+xy\le\sqrt{1-x^2}\cdot\sqrt{1-y^2}\\\iff\ &(1-x)(1-y)\le\sqrt{1-x^2}\sqrt{1-y^2},\end{aligned}$$

其中对任意的 $x\in[0,1]$, 都有 $1-x\le1-x^2\le\sqrt{1-x^2}$, 故该不等式成立; 接下来, 由 $(1-x)(1-y)\ge0$ 得

$$x+y-1\le xy;$$

最后, 由 $xy\le x, xy\le y$ 得

$$xy\le\min\{x,y\}.$$

结合以上不等式可得 $T_{\cos}(x,y) \leq T_L(x,y) \leq T_P(x,y) \leq T_M(x,y)$. □

在得到了命题1.8之后, 为了方便, 我们作出 t 模的图像, 如图3所示. 从图3可以直观地看出, 这四个 t 模应该是有大小关系的. $T_{\cos}$ 应该是它们之中最小的, 而 T_M 应该是最大的. 在后面的研究当中, 这几种 t 模都非常有用.

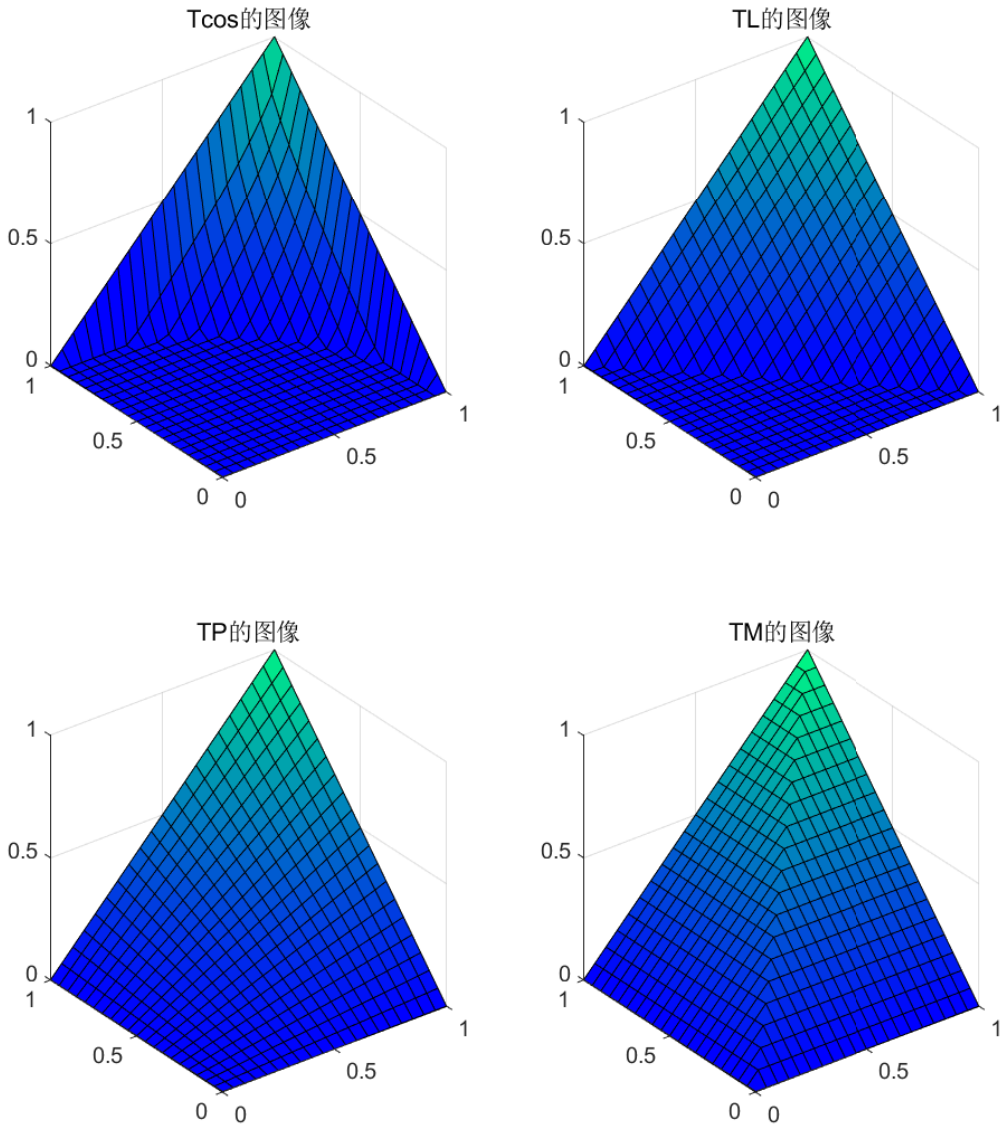


图 3: $T_{\cos}$、T_L、T_P 和 T_M 的图像

基于 t 模 T, 定义1.9探讨了 T 的逆运算.

定义 1.9 (蕴含与双蕴含)**.** 设 $T:[0,1]^2\to[0,1]$ 是左连续的 t 模, 称

$$\overrightarrow{T}(x,y)=\sup\{z\in[0,1]|T(x,z)\le y\},$$

为 T 的**蕴含**, 并称 $\overleftrightarrow{T}(x,y)=\min\left\{\overrightarrow{T}(x,y),\overrightarrow{T}(y,x)\right\}$ 为 T 的**双蕴含**.

例 1.10 (几种 t 模的蕴含)**.** 对于 $T_{\cos}(x,y)=\max\left\{xy-\sqrt{1-x^2}\cdot\sqrt{1-y^2},0\right\}$, 有

$$\overrightarrow{T_{\cos}}(x,y)=\begin{cases}\cos(\arccos y-\arccos x), & x>y,\\ 1, & \text{其他情况,}\end{cases}$$

对于 Lukasiewcz t 模 $T_L(x,y)=\max\{x+y-1,0\}$, 有

$$\overrightarrow{T_L}(x,y)=\min\{y-x+1,1\},\quad \overleftrightarrow{T_L}(x,y)=1-|x-y|;$$

对于乘积 t 模 $T_P(x,y)=xy$, 有

$$\overrightarrow{T_P}(x,y)=\begin{cases}\dfrac{y}{x}, & x>y,\\ 1, & \text{其他情况,}\end{cases}\qquad \overleftrightarrow{T_P}(x,y)=\min\left\{\frac{x}{y},\frac{y}{x}\right\};$$

对于 $T_M(x,y)=\min\{x,y\}$, 有

$$\overrightarrow{T_M}(x,y)=\begin{cases}y, & x>y,\\ 1, & \text{其他情况,}\end{cases}\qquad \overleftrightarrow{T_M}(x,y)=\min\{x,y\}.$$

在后面, t 模的双蕴含也是非常有用的.

1.3 T-等价关系及其构造

在 t 模的基础上, 我们给出 T-等价关系的概念, 见定义1.11.

定义 1.11 (T-等价关系)**.** 设 $T:[0,1]^2\to[0,1]$ 是 t 模, 函数 $E:\mathscr{X}\times\mathscr{X}\to[0,1]$ 满足

(1) 自反性, 也即对任意的 $x\in\mathscr{X}$, 有 $E(x,x)=1$;

(2) 对称性, 也即对任意的 $x,y\in\mathscr{X}$, 有 $E(x,y)=E(y,x)$;

(3) T-传递性, 也即对任意的 $x,y,z\in\mathscr{X}$, 有 $T(E(x,y),E(y,z))\le E(x,z)$,

则称 E 是 $\mathscr{X}$ 上的 T**-等价关系**.

基于命题1.8, 我们会发现对于不同的 T, T-等价关系之间也会有关系, 见命题1.12.

命题 1.12. *(1) 设 E 是 T_L-等价关系, 则 E 是 $T_{\cos}$-等价关系;*

(2) 设 E 是 T_P-等价关系, 则 E 也是 T_L 等价关系和 $T_{\cos}$-等价关系;

(3) 设 E 是 T_M-等价关系, 则 E 也是 T_P-等价关系、T_L-等价关系和 $T_{\cos}$-等价关系.

证明. 仅证明 (1), 其余的部分也可以类似证明: 设 E 是 T_L-等价关系, 则 E 具有自反性、对称性和 T_L-传递性, 根据命题1.8得

$$T_{\cos}(E(x,y),E(y,z)) \leq T_L(E(x,y),E(y,z)) \leq E(x,z),$$

从而 E 具有 $T_{\cos}$-传递性, 从而是 $T_{\cos}$-等价关系. □

上面的命题较为简单, 但是说明了, $T_{\cos}$-等价关系是最“弱”的等价关系, 而 T_M-等价关系是最强的等价关系. 为了方便理解, 作出关系图如图4所示.

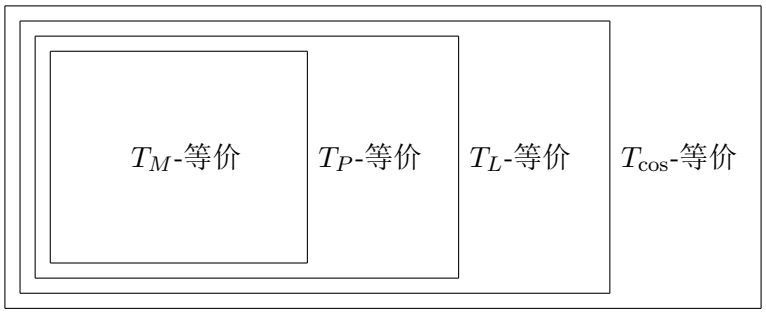


图 4: $T_{\cos}$-等价关系、T_L-等价关系、T_P-等价关系、T_M-等价关系之间的关系

接下来, 考虑 Archimedes 模对应的 T-等价关系, 其与伪距离之间有密切的关系. 这个关系由定理1.13给出.

定理 1.13. *设 $f:[0,1]\to[0,\infty]$ 是严格递增函数, $f(1)=0$, 定义 Archimedes 模*

$$T(x,y)=f^{-1}\left(\min\{f(x)+f(y),f(0)\}\right).$$

(1) 如果 d 是 $\mathscr{X}$ 上的伪距离, 则

$$E_d(x,y)=f^{-1}\left(\min\{d(x,y),f(0)\}\right)$$

是 $\mathscr{X}$ 上的 T-等价关系;

(2) 如果 E 是 $\mathscr{X}$ 上的 T-等价关系, 则函数

$$d_E(x,y)=f(E(x,y))$$

是 $\mathscr{X}$ 上的伪距离.

证明. 见参考文献 [5]、[6]. □

另外, T-等价关系也可以由 T 的双蕴含 $\overleftrightarrow{T}$ 构造, 如定理1.14所示.

定理 1.14. *设 $T:[0,1]^2\to[0,1]$ 是左连续的 t 模, $\mu_i:\mathscr{X}\to[0,1]$, 其中 $i\in I\neq\varnothing$, 则*

$$E(x,y)=\inf_{i\in I}\overleftrightarrow{T}\left(\mu_i(x),\mu_i(y)\right),\quad \forall x,y\in\mathscr{X}$$

是 $\mathscr{X}$ 上的 T-等价关系.

证明. 见参考文献 [7]、[8]、[9]. □

2 理论: 核函数与 T-等价关系之间的关系

接下来, 作为本篇报告的理论核心, 我们指出参考文献 [1] 和 [2] 中所说明的核函数与 T-等价关系之间的关系. 在本节中, 我们应用到的 t 模包含 $T_{\cos}$、T_L、T_p 和 T_M.

2.1 核函数是 $T_{\cos}$-等价关系

事实上, 对于满足自反性的核函数, 我们可以找到某一个 t 模 T, 使得它是 T-等价的. 更具体地说, 这个 t 模可以是我们上一节中所提到的 $T_{\cos}$.

定理 2.1 (Bernhard Moser, 2006). 设 $k : \mathscr{X} \times \mathscr{X} \to [0,1]$ 是 $\mathscr{X}$ 上的核函数, 且对任意的 $x \in \mathscr{X}$, $k(x,x) = 1$, 则 k 是 $T_{\cos}$-等价关系.

证明. 根据定义1.1知 k 具有自反性和对称性, 只需证明 k 还具有 $T_{\cos}$ 传递性, 也即对任意的 $x, y, z \in \mathscr{X}$, 有

$$T_{\cos}(k(x,y), k(y,z)) = \max\left\{k(x,y)k(y,z) - \sqrt{1-k^2(x,y)} \cdot \sqrt{1-k^2(y,z)}\right\} \leq k(x,z).$$

其中 $k(x,z) \geq 0$, 从而只需证明

$$\begin{aligned} & k(x,y)k(y,z) - \sqrt{1-k^2(x,y)} \cdot \sqrt{1-k^2(y,z)} \leq k(x,z) \\ \iff\ & k(x,y)k(y,z) - k(x,z) \leq \sqrt{1-k^2(x,y)} \cdot \sqrt{1-k^2(y,z)} \\ \iff\ & 1 + 2 \cdot k(x,y)k(y,z)k(z,x) - k^2(x,y) - k^2(y,z) - k^2(z,x) \geq 0. \end{aligned}$$

为了证明该不等式, 考虑矩阵

$$\boldsymbol{M} = \begin{bmatrix} 1 & k(x,y) & k(x,z) \\ k(x,y) & 1 & k(y,z) \\ k(x,z) & k(y,z) & 1 \end{bmatrix},$$

由 k 是核函数知 $\boldsymbol{M}$ 是正定矩阵, 从而

$$\det \boldsymbol{M} = 1 + 2 \cdot k(x,y)k(y,z)k(z,x) - k^2(x,y) - k^2(y,z) - k^2(z,x) \geq 0,$$

这便说明了原不等式成立, 从而 k 是 $T_{\cos}$-等价关系. □

定理2.1对核函数的要求只有两条, 一个是取值介于 0 和 1 之间, 另一个是具有自反性. 在 $\mathbb{R}^n$ 上, 我们可以对例1.2中的核函数进行改造, 并将其作为 $T_{\cos}$-等价关系, 应用到模糊聚类当中. 另外, 定理2.1可以导出定理2.2, 这说明了所有的满足上述条件的核都可以用类似定理1.14的公式来表示.

定理 2.2. *设 $k : \mathscr{X} \times \mathscr{X} \to [0,1]$ 是 $\mathscr{X}$ 上的核函数, 且对任意的 $x \in \mathscr{X}$, $k(x,x) = 1$, 则存在 $\mu_i : \mathscr{X} \to [0,1]$, 其中 $i \in I \neq \varnothing$, 及 t 模 T, 使得*

$$k(x,y) = \inf_{i \in I} \overleftrightarrow{T}(\mu_i(x), \mu_i(y)), \quad \forall x, y \in \mathscr{X}.$$

证明. 取 $I = \mathscr{X}$, 且对任意的 $x_0 \in \mathscr{X}$, 取 $\mu_{x_0}(x) = k(x, x_0)$. 并记

$$h(x,y) = \inf_{x_0 \in \mathscr{X}} \overleftrightarrow{T_{\cos}}(\mu_{x_0}(x), \mu_{x_0}(y)) = \inf_{x_0 \in \mathscr{X}} \overleftrightarrow{T_{\cos}}(k(x_0,x), k(x_0,y)),$$

接下来我们说明 $h(x,y) = k(x,y)$. 一方面, 根据定理2.1知核 $k(x,y)$ 一定是 $T_{\cos}$-等价关系, 从而具有 $T_{\cos}$-传递性, 因此

$$T_{\cos}(k(x_0,x), k(x_0,y)) \leq k(x,y) \implies h(x,y) \leq k(x,y);$$

另外一方面, 有

$$\begin{cases} T_{\cos}(k(x,y), k(x_0,y)) \leq k(x_0,x) \iff k(x,y) \leq \overrightarrow{T_{\cos}}(k(x_0,y), k(x_0,x)), \\ T_{\cos}(k(x,y), k(x_0,x)) \leq k(x_0,y) \iff k(x,y) \leq \overrightarrow{T_{\cos}}(k(x_0,x), k(x_0,y)), \end{cases}$$

因此

$$k(x,y) \leq \overleftrightarrow{T_{\cos}}(k(x_0,x), k(x_0,y)) \implies k(x,y) \leq h(x,y).$$

从而 $h(x,y) = k(x,y)$. □

2.2 T_M-等价关系是核函数

从本小节开始, 我们指出, 某些 t 模所生成的 T-等价关系是半正定的, 从而也是核函数. 首先考虑 T_M, 为了利用 T_M 来生成核函数, 我们还需要一些准备工作.

引理 2.3. *设 $k : \mathscr{X} \times \mathscr{X} \to [0,1]$, 则 k 是 T_M-等价关系当且仅当对任意的 $\alpha \in [0,1]$, k 的 α-水平截集 $[k]_\alpha$ 都是 $\mathscr{X}$ 上的等价关系.*

证明. ($\implies$) 设 $k : \mathscr{X} \times \mathscr{X} \to [0,1]$ 是 T_M-等价关系, 首先根据定义知 k 具有自反性、对称性和 T_M-传递性. 对任意的 $\alpha \in [0,1]$, 考虑集合 $[k]_\alpha$, 容易验证其满足自反性和对称性. 设 $(x,y), (y,z) \in [k]_\alpha$, 则根据 T_M-传递性得

$$k(x,y) \geq \alpha, k(y,z) \geq \alpha \implies T(x,z) \geq \min\{k(x,y), k(y,z)\} \geq \alpha,$$

从而 $(x,z)\in[k]_\alpha$, 这便证明了 $[k]_\alpha$ 具有传递性, 从而 $[k]_\alpha$ 是 $\mathscr{X}$ 上的等价关系.

($\Longleftarrow$) 设 $k:\mathscr{X}\times\mathscr{X}\to[0,1]$, 且对任意的 $\alpha\in[0,1]$, $[k]_\alpha$ 都是 $\mathscr{X}$ 上的等价关系. 首先, 由 $[k]_1$ 的自反性得 $k(x,x)=1$; 接下来, 对任意的 (x,y), 由 $[k]_{k(x,y)}$ 的对称性得 $k(y,x)\geq k(x,y)$, 同理也有 $k(x,y)\geq k(y,x)$, 从而 $k(x,y)=k(y,x)$; 最后, 对任意的 $(x,y),(y,z)$, 设 $\alpha=\min\{k(x,y),k(y,z)\}$, 由 $[k]_\alpha$ 的传递性得

$$(x,z)\in[k]_\alpha \implies k(x,z)\geq\alpha=\min\{k(x,y),k(y,z)\},$$

从而 $k(x,y)$ 是 T_M-等价关系. □

引理2.3给出了 T_M-等价关系的表示方式. 在此基础上, 我们来说明 T_M-等价关系都是核函数, 这对于我们后续的理论十分重要.

定理 2.4. *设 $k:\mathscr{X}\times\mathscr{X}\to[0,1]$ 是 T_M-等价关系, 则 k 是半正定的, 从而是核函数.*

证明. 对任意的 $x_1,x_2,\cdots,x_n\in\mathscr{X}$, 考虑集合

$$\{k(x_i,x_j)|1\leq i,j\leq n\}=\{\alpha_1,\alpha_2,\cdots,\alpha_m\},$$

其中 $0\leq\alpha_1\leq\cdots\leq 1$, 则有

$$k(x_i,x_j)=\alpha_1 I_{[k]_{\alpha_1}}(x_i,x_j)+\sum_{l=2}^{m}(\alpha_l-\alpha_{l-1})I_{[k]_{\alpha_l}}(x_i,x_j).$$

根据上式及引理2.3, k 在集合 $\{x_1,x_2,\cdots,x_n\}\times\{x_1,x_2,\cdots,x_n\}$ 上是等价关系的线性组合, 且系数非负, 从而 k 是半正定的. □

根据命题1.12, 我们知道 T_M-等价关系是最“强”的等价关系, 也即 T_M-等价关系一定是 $T_{\cos}$-等价关系、T_L-等价关系和 T_P-等价关系. 而定理2.1说明了, 满足自反性的核函数一定是 $T_{\cos}$-等价关系. 整理以上逻辑, 可得关系图如图5所示.

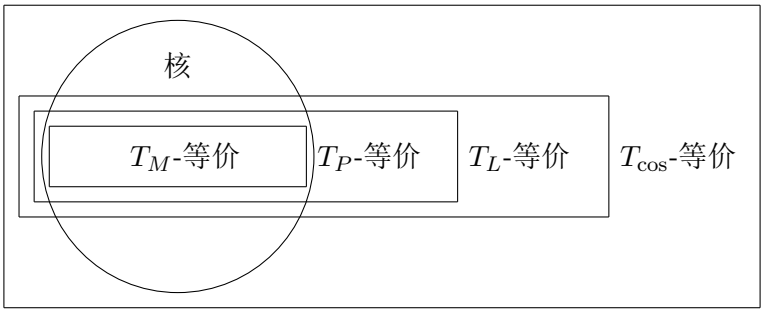


图 5: 核函数、$T_{\cos}$-等价关系、T_L-等价关系、T_P-等价关系、T_M-等价关系之间的关系

另外, 借助定理1.14, 可以进一步拓宽利用 T_M 所能构造的核函数.

推论. 设 $\mu_i : \mathscr{X} \to [0,1]$, 其中 $i \in I \neq \varnothing$, 则关系

$$E_M(x,y) = \inf_{i\in I} T_M(\mu_i(x), \mu_i(y)), \quad \forall x, y \in \mathscr{X}$$

是半正定的, 从而是核函数.

证明. 首先证明

$$E'_M(x,y) = \inf_{i\in I} \overleftrightarrow{T_M}(\mu_i(x), \mu_i(y)), \quad \forall x, y \in \mathscr{X}.$$

是半正定的. 根据定理1.14知 $E'_M(x,y)$ 是 T_M-等价关系, 再根据定理2.4知 T_M-等价关系是半正定的, 从而 $E'_M(x,y)$ 是半正定的, 从而是核函数. 再根据例1.10知

$$\overleftrightarrow{T_M}(x,y) = \min\{x,y\} = T_M(x,y), \quad \forall x, y \in [0,1],$$

从而 $E_M = E'_M$, 这便说明了 E_M 是半正定的, 从而是核函数. □

2.3 用 T_L 和 T_P 生成核函数

接下来, 考虑 T_L 和 T_P, 我们来说明了这两种 t-模也可以生成半正定的等价关系, 从而也是核函数. 在这里, 我们便需要应用到 T_L 和 T_P 的双蕴含, 而我们在例1.10中已经得到了它们的双蕴含的表达式. 后面, 我们将直接应用该结果.

定理 2.5. 设 $v : \mathscr{X} \to [0,1]$, $h : [0,1] \to [0,1]$ 是单位区间上的同构, 且是 $\mathbb{C}$ 上形如 $f(x) = \sum_{k\geq 0} c_k x^k$ 的整函数在 $\mathbb{R}$ 上的限制, 其中对所有的 $k \geq 0$ 有 $c_k \geq 0$.

(1) 对于 *Lukasiewcz* t 模 T_L, 关系

$$E_{L,h}(x,y) = h\left(\overleftrightarrow{T_L}\left(h^{-1}(v(x)), h^{-1}(v(y))\right)\right)$$

是半正定的, 从而是核函数;

(2) 对于乘积 t 模 T_P, 关系

$$E_{P,h}(x,y) = h\left(\overleftrightarrow{T_P}\left(h^{-1}(v(x)), h^{-1}(v(y))\right)\right)$$

是半正定的, 从而是核函数.

证明. 为了方便, 记 $\mu_i = h^{-1}(v(x_i))$. 不失一般性地, 设 $\{\mu_i\}$ 是单调递减的.

(1) 根据例1.10, $\overleftrightarrow{T_L}(x,y) = \min\{\overrightarrow{T_L}(x,y), \overrightarrow{T_L}(y,x)\} = 1 - |x-y|$. 根据定理1.3, 要证

明 $E_{P,h}$ 是半正定的, 只需证明对任意的 $n \geq 1$, 都有

$$D_n = \begin{vmatrix} 1 & 1-(\mu_1-\mu_2) & \cdots & 1-(\mu_1-\mu_n) \\ 1-(\mu_1-\mu_2) & 1 & \cdots & 1-(\mu_2-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ 1-(\mu_1-\mu_n) & 1-(\mu_2-\mu_n) & \cdots & 1 \end{vmatrix} \geq 0.$$

不断进行列变换, 计算得

$$\begin{aligned} D_n &= \begin{vmatrix} 1 & \mu_2-\mu_1 & \cdots & \mu_n-\mu_{n-1} \\ 1-(\mu_1-\mu_2) & \mu_1-\mu_2 & \cdots & \mu_n-\mu_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1-(\mu_1-\mu_n) & \mu_1-\mu_2 & \cdots & \mu_{n-1}-\mu_n \end{vmatrix} \\ &= \prod_{i=2}^{n}(\mu_{i-1}-\mu_i) \cdot \begin{vmatrix} 1 & -1 & \cdots & -1 \\ 1-(\mu_1-\mu_2) & 1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ 1-(\mu_1-\mu_n) & 1 & \cdots & 1 \end{vmatrix} \\ &= \prod_{i=2}^{n}(\mu_{i-1}-\mu_i) \cdot (2-(\mu_1-\mu_n)) \cdot 2^{n-2} \\ &\geq 0, \end{aligned}$$

从而 $E_{L,h}$ 是半正定的.

(2) 根据例1.10,

$$\begin{aligned} \overleftrightarrow{T_P}(x,y) &= \min\{\overrightarrow{T_P}(x,y), \overrightarrow{T_P}(y,x)\} \\ &= \begin{cases} \min\left\{\dfrac{b}{a}, \dfrac{a}{b}\right\}, & \text{如果 } a,b>0, \\ 1, & \text{如果 } a=b=0, \\ 0, & \text{其他情况.} \end{cases} \end{aligned}$$

在假设 $\{\mu_i\}$ 单调递增的基础上, 不妨再假设自第 i_0 项开始 $\mu_i = 0$, 则当 $i < i_0, j \geq i_0$ 时, $\overleftrightarrow{T_P}(\mu_i, \mu_j) = 0$; 而当 $i, j \geq i_0$ 时, $\overleftrightarrow{T_P}(\mu_i, \mu_j) = 1$. 根据分块矩阵的行列式计算法则, 只要行列式的左上角大于等于零, 整个行列式也是大于等于零的. 从而, 不妨假设 $\{\mu_i\}$ 非零.

根据定理1.3, 结合以上假设, 只需证明对任意的 $n \geq 1$, 都有

$$D_n = \begin{vmatrix} 1 & \frac{\mu_2}{\mu_1} & \cdots & \frac{\mu_n}{\mu_1} \\ \frac{\mu_2}{\mu_1} & 1 & \cdots & \frac{\mu_n}{\mu_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mu_n}{\mu_1} & \frac{\mu_n}{\mu_2} & \cdots & 1 \end{vmatrix} \geq 0.$$

将第 i 列乘上 $-\frac{\mu_{i+1}}{\mu_i}$, 并加到第 $i+1$ 列, 计算得

$$\begin{aligned} D_n &= \begin{vmatrix} 1 & 0 & \cdots & 0 \\ * & 1-\left(\frac{\mu_2}{\mu_1}\right)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 1-\left(\frac{\mu_n}{\mu_{n-1}}\right) \end{vmatrix} \\ &= \prod_{i=1}^{n-1}\left\{1-\left(\frac{\mu_{i+1}}{\mu_i}\right)^2\right\} \geq 0, \end{aligned}$$

从而 $E_{P,h}$ 是半正定的. □

定理 2.6. 设 $\mu_i : \mathscr{X} \to [0,1]$, 其中 $i \in I \neq \varnothing$, $\lambda_i \in [0,1], 1 \leq i \leq n$, 且 $\sum_{i=1}^{n} \lambda_i = 1$.

(1) 对于 *Lukasiewcz* t 模 T_L, 关系

$$\widetilde{E_L}(x,y) = \sum_{i=1}^{n} \lambda_i \overleftrightarrow{T_L}(\mu_i(x), \mu_i(y))$$

是 T_L-等价关系, 也是核函数;

(2) 对于乘积 t 模 T_P, 关系

$$\widetilde{E_P}(x,y) = \prod_{i=1}^{n} \left(\overleftrightarrow{T_P}(\mu_i(x), \mu_i(y))\right)^{\lambda_i}$$

是 T_P-等价关系, 也是核函数.

证明. (1) 容易验证 $\widetilde{E_L}$ 满足 T_L-传递性, 从而是 T_L-等价关系. 根据定理1.3及定理2.5, 知 $\widetilde{E_L}$ 是半正定的, 从而是核函数.

(2) 容易验证 $\widetilde{E_P}$ 满足 T_P-传递性, 从而是 T_P-等价关系. 根据定理1.3及定理2.5, 知 $\widetilde{E_L}$ 是半正定的, 从而是核函数. □

2.4 T-等价关系的生成方法综述

在前面几个小节中, 首先为了探讨如何用核函数生成等价关系, 我们证明了所有的可以作为等价关系的核函数都是 $T_{\cos}$-等价的 (定理2.1), 并且在此基础上, 说明了任何的核函数都可以用 t 模的双蕴含来表示 (定理2.2). 接下来, 为了用等价关系生成核函数, 我们考虑 t 模 T_M, 在 T_M 的性质 (引理2.3) 的基础上, 证明了 T_M-等价关系都是核函数 (定理2.4); 我们再考虑 t 模 T_L 和 T_P, 给出了用 T_L 和 T_P 生成核函数的方法 (定理2.5), 以及用用 T_L 和 T_P 生成 T-等价关系的方法 (定理2.6).

为了方便应用于后续的动态模糊聚类中, 在此设 $\mathscr{X}=\mathbb{R}^n$, $\boldsymbol{x},\boldsymbol{y}\in\mathscr{X}$ 是 n 维实向量, 整理 T-等价关系的生成方法. 首先是借助核函数来生成等价关系.

(1) 首先, 设 $k:\mathbb{R}^n\times\mathbb{R}^n\to[0,1]$ 是核函数, 且满足 $k(\boldsymbol{x},\boldsymbol{x})=1$, 则其是 $T_{\cos}$-等价关系. 对于例1.2给出的 Gauss 核函数

$$k(\boldsymbol{x},\boldsymbol{y})=\exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\},$$

容易验证 $k(\boldsymbol{x},\boldsymbol{x})=1$, 从而其是 $T_{\cos}$-等价关系; 而对于表1中的核函数, 也都有 $k(\boldsymbol{x},\boldsymbol{x})=1$, 并且 Laplace 核函数、二次有理核函数和逆多元二次核函数 (当 $c=1$ 时) 取值都介于 0 和 1 之间, 从而它们都是 $T_{\cos}$-等价关系.

(2) 在 (1) 的基础上, 对于不满足 $k(\boldsymbol{x},\boldsymbol{x})=1$ 的函数, 也可以尝试将其单位化, 例如对于例1.2给出的线性核函数, 将其改写为

$$k(\boldsymbol{x},\boldsymbol{y})=\frac{|\boldsymbol{x}^T\boldsymbol{y}|}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|}=\frac{|(\boldsymbol{x},\boldsymbol{y})|}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|}=|\cos\langle\boldsymbol{x},\boldsymbol{y}\rangle|,$$

容易验证 $k(\boldsymbol{x},\boldsymbol{x})=1$, 从而其是 $T_{\cos}$-等价关系.

除了应用已有的核函数以外, 我们也可以用 t 模来生成等价关系.

(1) 设 $T:[0,1]^2\to[0,1]$ 是左连续的 t 模, $\mu_i:\mathbb{R}\to[0,1]$, 其中 $i\in I\neq\varnothing$, 根据定理1.14知关系

$$E(x,y)=\inf_{i\in I}\overleftrightarrow{T}(\mu_i(x),\mu_i(y))$$

是 T-等价关系, 其中 T 可以是 $T_{\cos}$、T_L、T_P 和 T_M.

(2) 在 (1) 的基础上, 根据定理2.5, 设 T 是 T_L 或 T_P, $v:\mathbb{R}\to[0,1]$, $h:[0,1]\to[0,1]$ 是单位区间上的同构, 且是 $\mathbb{C}$ 上形如 $f(x)=\sum_{k\geq0}c_kx^k$ 的整函数在 $\mathbb{R}$ 上的限制, 其中对所有的 $k\geq0$ 有 $c_k\geq0$, 则关系

$$E(x,y)=h\left(\overleftrightarrow{T}\left(h^{-1}(v(x)),h^{-1}(v(y))\right)\right)$$

是核函数, 从而是 $T_{\cos}$-等价关系.

(3) 在 (1) 和 (2) 的基础上, 根据定理2.6, 设 $T = T_L$, 则关系

$$\widetilde{E_L}(x,y) = \sum_{i=1}^{n} \lambda_i \overleftrightarrow{T_L}(\mu_i(x), \mu_i(y))$$

是 T_L-等价关系; 设 $T = T_P$, 则关系

$$\widetilde{E_P}(x,y) = \prod_{i=1}^{n} \left(\overleftrightarrow{T_P}(\mu_i(x), \mu_i(y)) \right)^{\lambda_i}$$

是 T_P-等价关系.

3 方法: 动态模糊聚类与结果的评价

3.1 动态模糊聚类方法概述

将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类. 抽象成数学语言后, 聚类的问题是这样: 设共有 n 个事物, $U = \{u_1, u_2, \cdots, u_n\}$ 为待分类的事物的全体, 每个事物都有 m 个特征, 对于第 j 个事物, 其特征记为 $u_j = (x_{j1}, x_{j2}, \cdots, x_{jm})$. 我们的目的是, 将 U 分为 c 个不同的类, 也即找到 U 的 c 个子集 $U_1, U_2, \cdots, U_c$, 使得

$$\bigcup_{i=1}^{c} U_i = U, \quad \text{且} \quad U_i \cap U_j = \varnothing, \quad i \neq j.$$

上面的每个集合 U_i 也被称为**簇**. 在不同的聚类的方法中, 动态模糊聚类是一种应用了模糊等价关系的聚类方法. 在介绍这种聚类方法之前, 还需要介绍一些概念.

定义 3.1 (等价矩阵、相似矩阵与传递矩阵)**.** 设 $\boldsymbol{R} = [r_{ij}]_{n\times n} \in [0,1]^{n\times n}$ 满足

(1) 自反性, 也即 $\boldsymbol{I} \subset \boldsymbol{R}$;

(2) 对称性, 也即 $\boldsymbol{R}^T = \boldsymbol{R}$;

(3) 传递性, 也即 $\boldsymbol{R}^2 \subset \boldsymbol{R}$,

则称 $\boldsymbol{R}$ 是**等价矩阵**. 若 $\boldsymbol{R}$ 仅满足 (1) 和 (2), 则称 $\boldsymbol{R}$ 是**相似矩阵**. 若 $\boldsymbol{R}$ 仅满足 (3), 则称 $\boldsymbol{R}$ 是**传递矩阵**.

定义 3.2 (λ-截矩阵)**.** 设 $\boldsymbol{R} = [r_{ij}]_{n\times n} \in [0,1]^{n\times n}$, $\lambda \in (0,1)$, 则矩阵

$$\boldsymbol{R}_\lambda = \left[r_{ij}^{(\lambda)} \right]_{n\times n} \quad \text{其中} \quad r_{ij}^{(\lambda)} = \begin{cases} 1, & r_{ij} \geq \lambda, \\ 0, & r_{ij} < \lambda \end{cases}$$

称为 $\boldsymbol{R}$ 的 **λ-截矩阵**.

在聚类时, 如果可以得到等价矩阵 $\boldsymbol{R}$, 再选取适当的 $\lambda \in (0,1)$, 得到 λ-截矩阵 $\boldsymbol{R}_\lambda$, 就可以根据 $\boldsymbol{R}_\lambda$ 所得到的关系对 U 进行分类. 然而, 在实际应用中, 得到模糊等价矩阵并

不容易, 通常所能得到的都是模糊相似矩阵, 也即满足自反性和对称性的矩阵. 因此, 我们需要研究从模糊相似矩阵构造模糊传递矩阵的方法.

定义 3.3 (传递闭包)**.** 设 $\boldsymbol{R}=[r_{ij}]_{n\times n}\in[0,1]^{n\times n}$, 包含 $\boldsymbol{R}$ 并被任何包含 $\boldsymbol{R}$ 的传递矩阵所包含的传递矩阵称为 $\boldsymbol{R}$ 的**传递闭包**, 记为 $t(\boldsymbol{R})$.

定理 3.4 (传递闭包的存在性)**.** 设 $\boldsymbol{R}=[r_{ij}]_{n\times n}\in[0,1]^{n\times n}$, 则

$$t(\boldsymbol{R})=\bigcup_{k=1}^{\infty}\boldsymbol{R}^k.$$

证明. 记 $\boldsymbol{A}=\bigcup\limits_{k=1}^{\infty}\boldsymbol{R}^k$. 首先有 $\boldsymbol{R}\subset\boldsymbol{A}$, 也即 $\boldsymbol{R}$ 包含 $\boldsymbol{A}$; 接下来, 根据

$$\boldsymbol{A}^2=\left(\bigcup_{k=1}^{\infty}\boldsymbol{R}^k\right)^2=\bigcup_{k=2}^{\infty}\boldsymbol{R}^k\subset\bigcup_{k=1}^{\infty}\boldsymbol{R}^k=\boldsymbol{A},$$

知 $\boldsymbol{A}$ 是传递矩阵; 最后, 对任意的包含 $\boldsymbol{R}$ 的传递矩阵 $\boldsymbol{Q}$, 都有

$$\boldsymbol{R}^k\subset\boldsymbol{Q}^k\subset\boldsymbol{Q}\implies\boldsymbol{A}=\bigcup_{k=1}^{\infty}\boldsymbol{R}^k\subset\boldsymbol{Q},$$

从而根据定义知 $t(\boldsymbol{R})=\boldsymbol{A}$. □

定理3.4在实际计算中不容易应用. 而定理3.5说明了, 实际上只要经过 n 次并运算, 就可以得到传递闭包.

定理 3.5 (传递闭包的计算)**.** 设 $\boldsymbol{R}=[r_{ij}]_{n\times n}\in[0,1]^{n\times n}$, 则

$$t(\boldsymbol{R})=\bigcup_{k=1}^{n}\boldsymbol{R}^k.$$

证明. 根据 $t(\boldsymbol{R})=\bigcup\limits_{k=1}^{\infty}\boldsymbol{R}^k$, 设 $(x,y)\in t(\boldsymbol{R})$, 则一定存在 k, 使得 $(x,y)\in\boldsymbol{R}^k$. 不妨设 k 是最小的自然数, 并设 G 是 $\boldsymbol{R}$ 的关系图, 一共有 n 个点.

一方面有 $k\leq n$. 假设 $k>n$, 则 $k\geq n+1$. 由 $(x,y)\in\boldsymbol{R}^k$ 知 G 上存在长度为 k 的路径, 连接 x 和 y. 该路径经过了 $k-1$ 个点, 超过了图 G 的点的个数, 从而一定有重复经过的点. 这便说明了, 可以找到更短的路径, 此与 k 是最小的自然数矛盾. 另外一方面, 对任何的 $(x,y)\in t(\boldsymbol{R})$, 都存在 $k\leq n$, 使得 $(x,y)\in\boldsymbol{R}^k$, 这便说明了 $t(\boldsymbol{R})=\bigcup\limits_{k=1}^{n}\boldsymbol{R}^k$ □

在以上理论的基础上, 动态模糊聚类遵循以下几个步骤:

(1) 对已有数据 $U=\{u_1,u_2,\cdots,u_n\}$ 进行标准化, 得到 $U'=\{u_1',u_2',\cdots,u_n'\}$;

(2) 构造相似矩阵 $\boldsymbol{R}=[r_{ij}]_{n\times n}\in[0,1]^{n\times n}$;

(3) 求传递闭包 $t(\boldsymbol{R}) = \bigcup_{k=1}^{n} \boldsymbol{R}^k$;

(4) 求传递闭包的 λ-截矩阵 $t(\boldsymbol{R})_\lambda$, 并根据该结果画图, 得到聚类结果.

在本篇报告中, 对第 k 个属性, 使用极差进行标准化, 也即令

$$x'_{jk} = \frac{x_{jk} - \min\limits_{1\le k\le m} x_{jk}}{\max\limits_{1\le k\le m} x_{jk} - \min\limits_{1\le k\le m} x_{jk}} \in [0,1], \quad \forall 1 \le k \le m,$$

并记标准化后的数据为 $U' = \{u'_1, u'_2, \cdots, u'_n\}$; 接下来, 设 $E : \mathbb{R}^m \times \mathbb{R}^m \to [0,1]$ 是 T-等价关系, 构造相似矩阵

$$r_{ij} = E(u'_i, u'_j), \quad \boldsymbol{R} = [r_{ij}]_{n\times n} \in [0,1]^{n\times n};$$

最后, 求出传递闭包 $t(\boldsymbol{R})$, 并根据其 λ-截矩阵得到聚类结果.

3.2 聚类结果评价方法概述

3.2.1 轮廓值

在机器学习与数据挖掘领域, **轮廓值**是一种反映数据聚类结果一致性的指标, 可以用于评估聚类后簇与簇之间的离散程度. 轮廓的取值范围为 $[-1,1]$. 如果某一样本的轮廓接近 1, 则说明样本聚类结果合理.

假设某一数据集 $U = \{u_1, u_2, \cdots, u_n\}$ 分成了 c 个簇 $U_1, U_2, \cdots, U_c$, 满足

$$\bigcup_{i=1}^{c} U_i = U, \quad \text{且} \quad U_i \cap U_j = \varnothing, \quad i \ne j.$$

设 d_{ij} 为样本 u_i 与 u_j 之间的距离, 并设样本 $u_j \in U_i$, 则计算样本 u_j 与其他样本之间的**平均距离**为

$$a_j = \frac{1}{|U_i| - 1} \sum_{k\in C_i, k\ne j} d_{jk}.$$

其中, 不计算样本与自身的距离 d_{jj}, 故计算平均值时样本总数为 $|U_i| - 1$. a_j 反映了样本 u_j 当前的聚类结果的优劣, 值越小, 聚类结果越好. 接下来, 对于样本 $u_j \in U_i$, 定义样本与某簇 U_k 的**平均相异性**为 u_j 到簇 U_k 内所有样本的距离均值, 其中 $U_k \ne U_i$. 记

$$b_j = \min_{k\ne i} \frac{1}{|U_k|} \sum_{j\in U_k} d_{ij},$$

为最小的平均相异性, 使得上式取最小值的簇 U_k 称为 u_i 的**相邻簇**.

结合上述内容, 设样本 $u_j \in U_i$, 并设 $|U_i| > 1$, 我们定义样本 u_j 的轮廓值为

$$s_j = \frac{b_j - a_j}{\max\{a_j, b_j\}} = \begin{cases} 1 - \dfrac{a_j}{b_j}, & \text{如果 } a_j < b_j, \\ 0, & \text{如果 } a_j = b_j, \\ \dfrac{b_j}{a_j} - 1, & \text{如果 } a_j > b_j. \end{cases}$$

对于上述定义, 显然有 $-1 \leq s_j \leq 1$. s_j 越接近 1, 说明聚类的结果越合理的.

3.2.2 相关性度量

为了评价我们的聚类结果的好坏, 我们还可以使用我们熟知的**相关系数**, 通过计算任两个样本之间的相关系数, 把他们放在一个矩阵 $\boldsymbol{A} = [a_{ij}]_{n\times n}$ 中, 其中 a_{ij} 表示 $u_i u_j$ 之间的相关系数. 我们将样本点的顺序作重新排列, 在我们的模糊聚类方法中被聚为一类的两个点被排列在相邻位置. 我们通过**热力图**表示相关性矩阵, 颜色越深表示相关系数越大, 这样如果热力图代表的相关性矩阵, 那么对角线上的子矩阵的颜色明显整体深于其周围颜色表示聚类效果较好.

4 结果: 基于核函数与 T-等价关系的动态模糊聚类

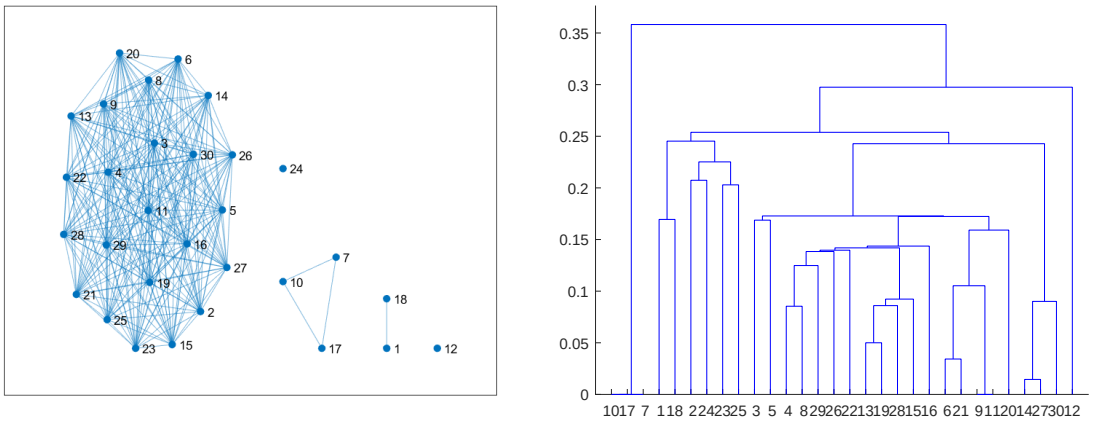
本数据来源于Kaggle, 是一组排球运动员的技术统计数据, 数据的特征包括每个运动员的发球得分数、救球数及扣球成功率等, 属于无标签数据. 我们将用基于不同的 T-等价关系的动态模糊聚类, 对该组数据进行处理, 观察数据在不同方法下得到的聚类结果.

4.1 基于线性核函数的聚类结果

本节中使用线性核函数

$$k(\boldsymbol{x},\boldsymbol{y})=\frac{|(\boldsymbol{x},\boldsymbol{y})|}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|},$$

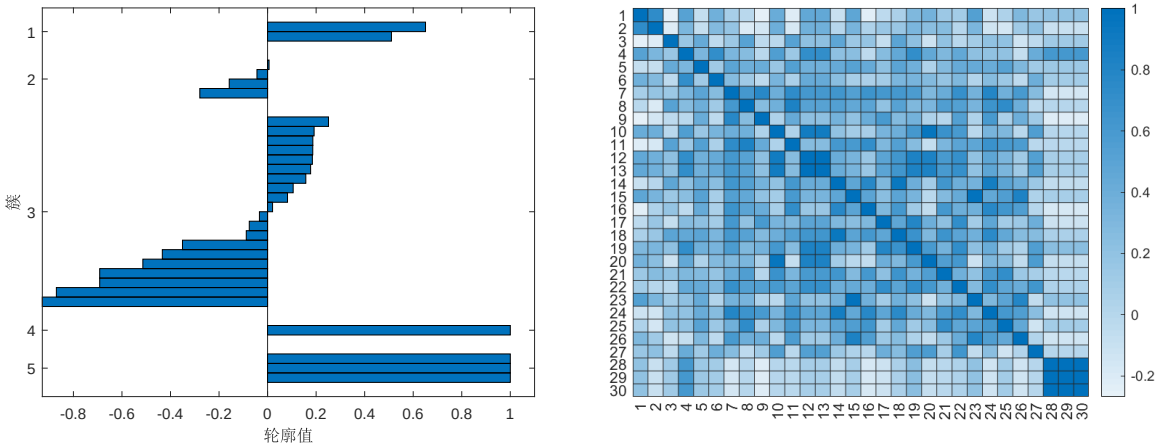
其是 $T_{\cos}$-等价关系. 根据传递闭包的 λ-截矩阵得到聚类结果如图6(a) 所示. 同时, 画出动态聚类图如图6(b) 所示.



(a) λ-截矩阵所得的聚类结果　　(b) 动态聚类图

图 6: 基于线性核函数的聚类结果

计算出每个点的轮廓值, 以及画出相关性度量的热力图如下图所示.



(a) 轮廓值　　(b) 热力图

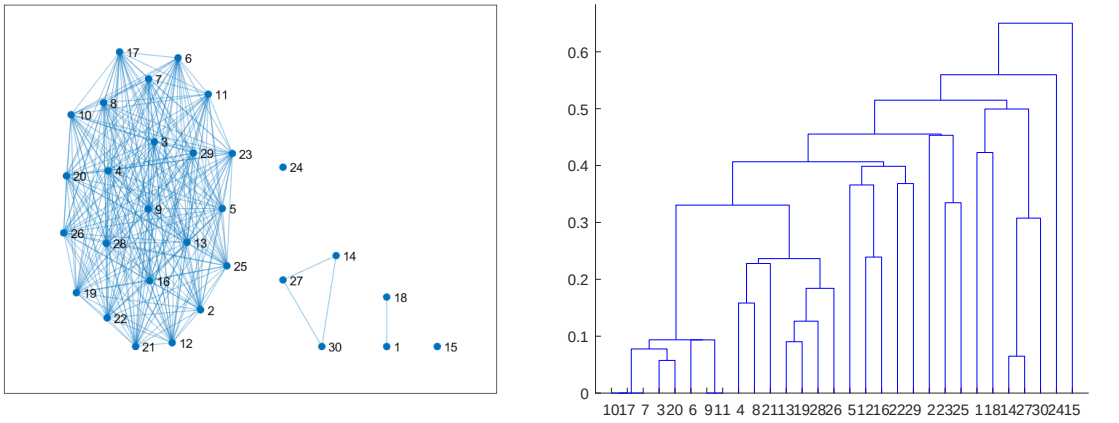
图 7: 基于线性核函数的聚类结果的评价

4.2 基于 Gauss 核函数的聚类结果

本节中使用 $\sigma = 1$ 的 Gauss 核函数

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right\},$$

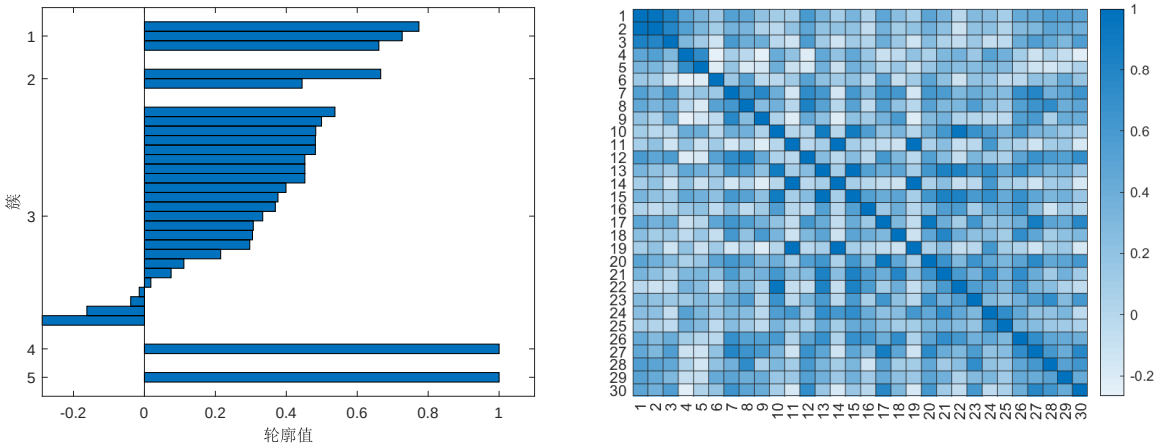
其是 $T_{\cos}$-等价关系. 根据传递闭包的 λ-截矩阵得到聚类结果如图8(a) 所示. 同时, 画出动态聚类图如图8(b) 所示.



(a) λ-截矩阵所得的聚类结果　　(b) 动态聚类图

图 8: 基于 Gauss 核函数的聚类结果

计算出每个点的轮廓值, 以及画出相关性度量的热力图如下图所示.



(a) 轮廓值　　(b) 热力图

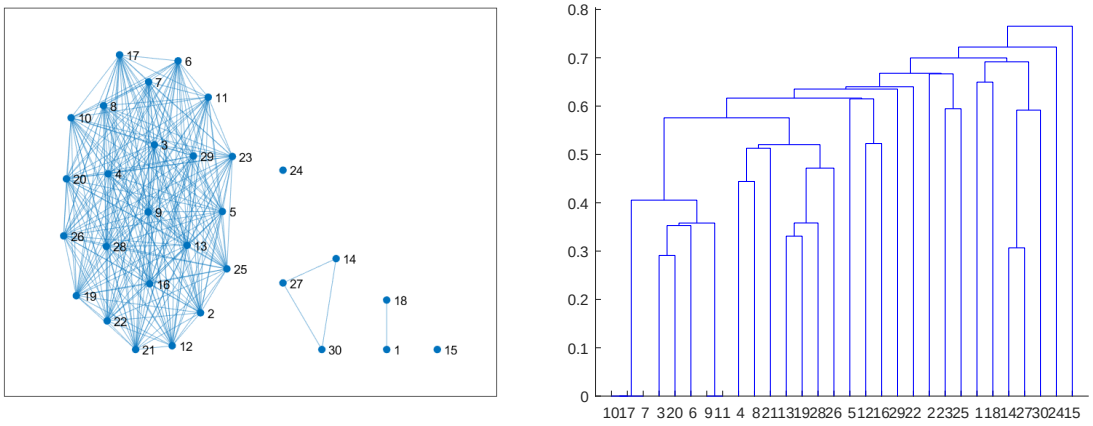
图 9: 基于 Gauss 核函数的聚类结果的评价

4.3 基于 Laplace 核函数的聚类结果

本节中使用 $\sigma = 1$ 的 Laplace 核函数

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sigma}\right\},$$

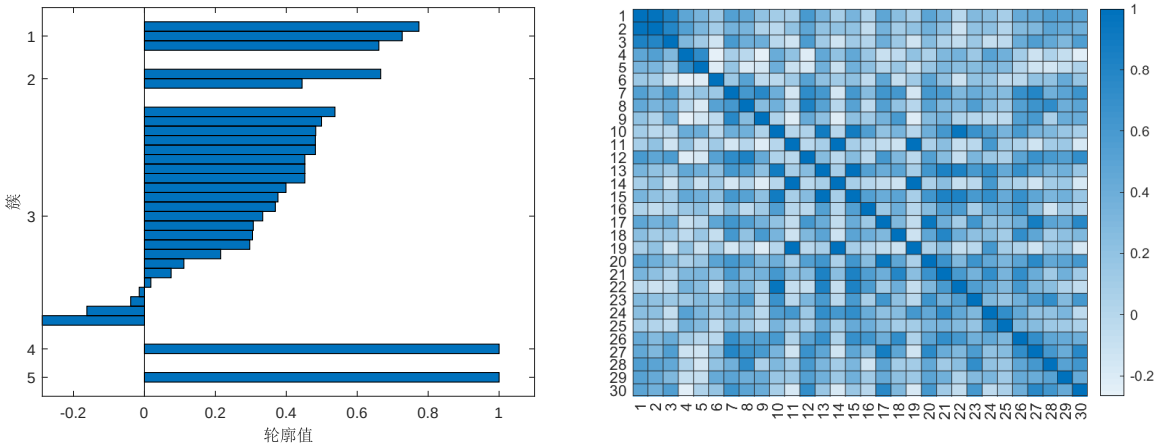
其是 $T_{\cos}$-等价关系. 根据传递闭包的 λ-截矩阵得到聚类结果如图10(a) 所示. 同时, 画出动态聚类图如图10(b) 所示.



(a) λ-截矩阵所得的聚类结果　　(b) 动态聚类图

图 10: 基于 Laplace 核函数的聚类结果

计算出每个点的轮廓值, 以及画出相关性度量的热力图如下图所示.



(a) 轮廓值　　(b) 热力图

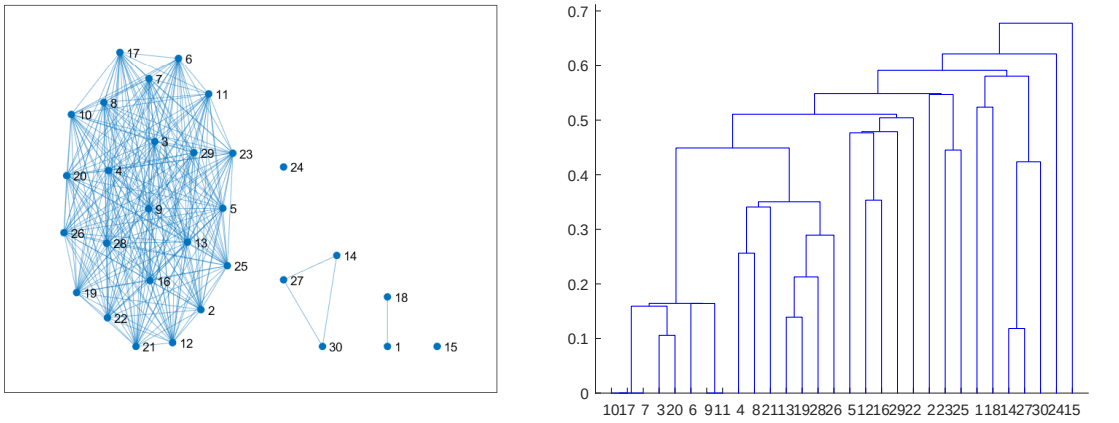
图 11: 基于 Laplace 核函数的聚类结果的评价

4.4 基于二次有理核函数的聚类结果

本节中使用 $c=1$ 的二次有理核函数

$$k(\boldsymbol{x},\boldsymbol{y})=1-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^2+c^2},$$

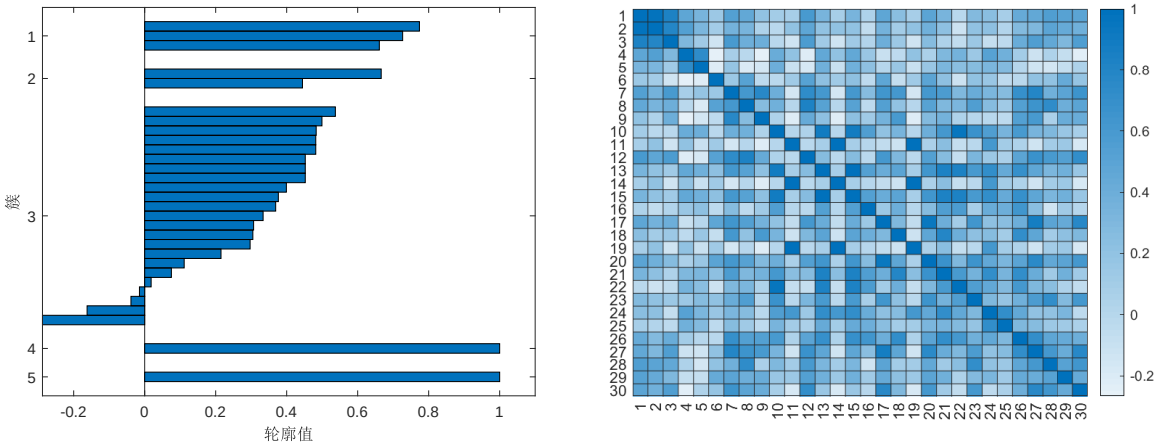
其是 $T_{\cos}$-等价关系. 根据传递闭包的 λ-截矩阵得到聚类结果如图12(a) 所示. 同时, 画出动态聚类图如图12(b) 所示.



(a) λ-截矩阵所得的聚类结果　　(b) 动态聚类图

图 12: 基于二次有理核函数的聚类结果

计算出每个点的轮廓值, 以及画出相关性度量的热力图如下图所示.



(a) 轮廓值　　(b) 热力图

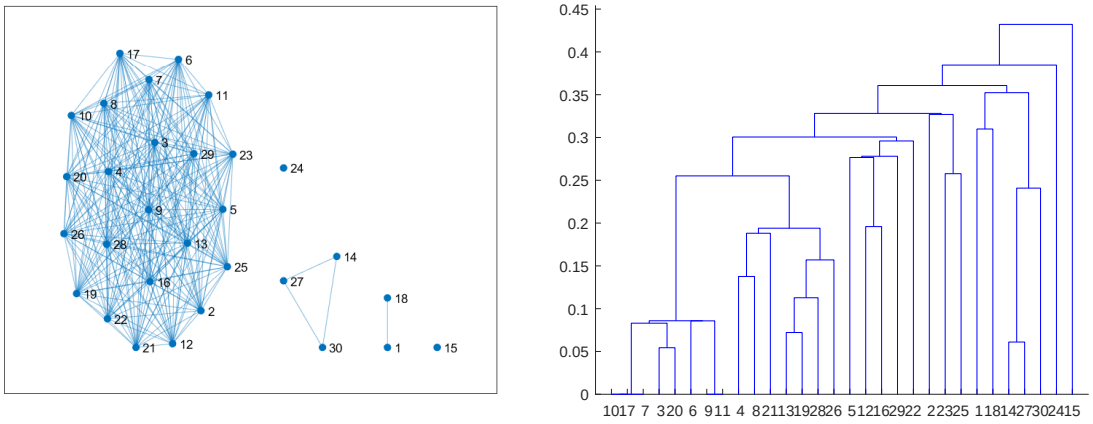
图 13: 基于二次有理核函数的聚类结果的评价

4.5 基于逆多元二次核函数的聚类结果

本节中使用 $c=1$ 的二次有理核函数

$$k(\boldsymbol{x},\boldsymbol{y})=\frac{1}{\sqrt{\|\boldsymbol{x}-\boldsymbol{y}\|^2+c^2}},$$

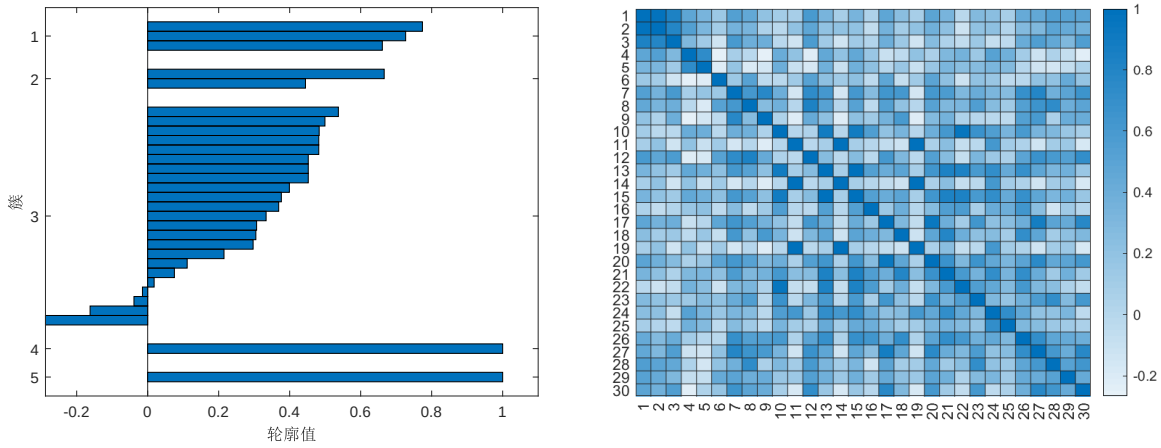
其是 $T_{\cos}$-等价关系. 根据传递闭包的 λ-截矩阵得到聚类结果如图14(a) 所示. 同时, 画出动态聚类图如图14(b) 所示.



(a) λ-截矩阵所得的聚类结果　　(b) 动态聚类图

图 14: 基于逆多元二次核函数的聚类结果

计算出每个点的轮廓值, 以及画出相关性度量的热力图如下图所示.



(a) 轮廓值　　(b) 热力图

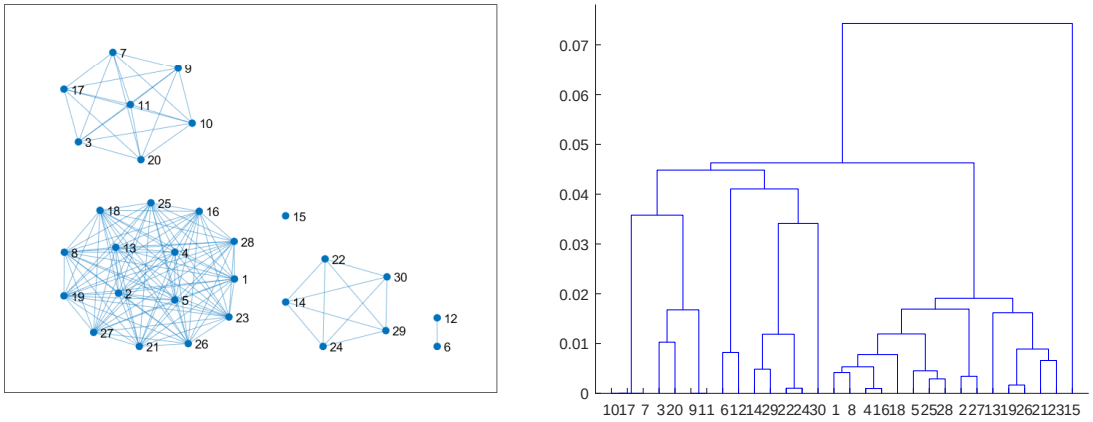
图 15: 基于逆多元二次核函数的聚类结果的评价

4.6 基于 T_L-等价关系的聚类结果

本节中取 $\mu(\boldsymbol{x}) = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, 并令

$$E(\boldsymbol{x}, \boldsymbol{y}) = \overleftrightarrow{T_L}(\mu(\boldsymbol{x}), \mu(\boldsymbol{y})),$$

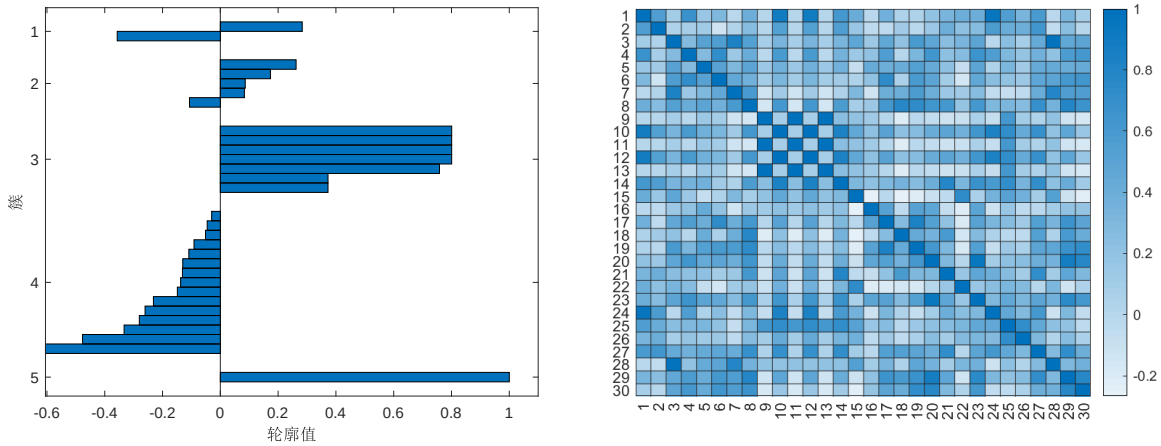
其是 T_P-等价关系. 根据传递闭包的 λ-截矩阵得到聚类结果如图16(a) 所示. 同时, 画出动态聚类图如图16(b) 所示.



(a) λ-截矩阵所得的聚类结果　　(b) 动态聚类图

图 16: 基于 T_L-等价关系的聚类结果

计算出每个点的轮廓值, 以及画出相关性度量的热力图如下图所示.



(a) 轮廓值　　(b) 热力图

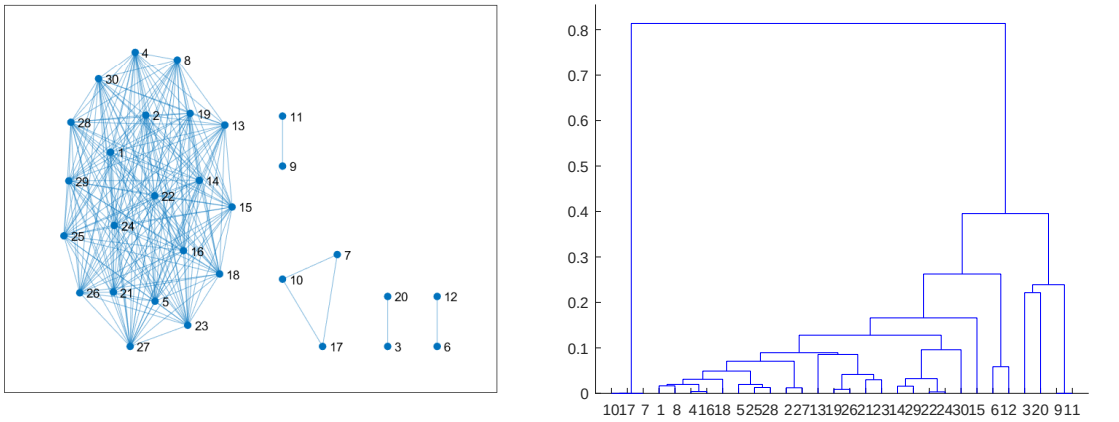
图 17: 基于 T_L-等价关系的聚类结果的评价

4.7 基于 T_P-等价关系的聚类结果

本节中取 $\mu(\boldsymbol{x}) = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, 并令

$$E(\boldsymbol{x}, \boldsymbol{y}) = \overleftrightarrow{T_P}(\mu(\boldsymbol{x}), \mu(\boldsymbol{y})),$$

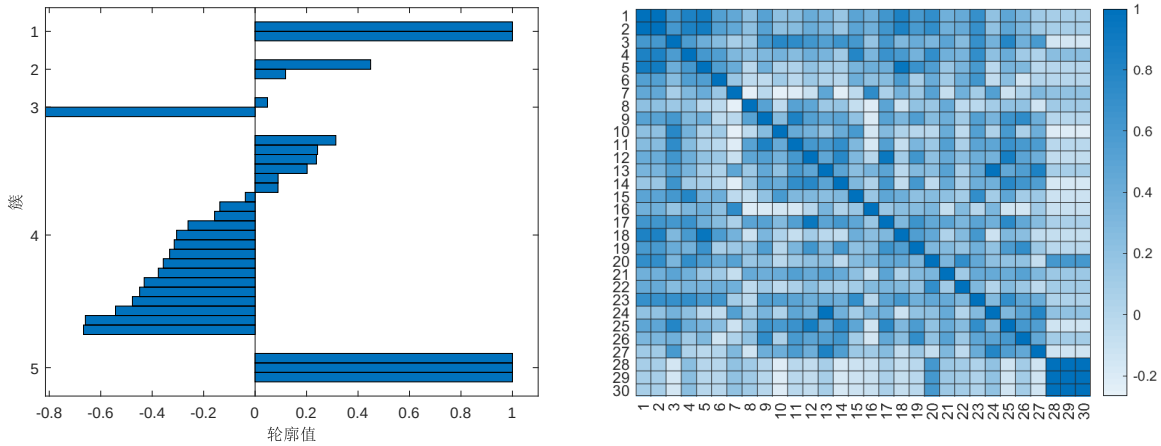
其是 T_P-等价关系. 根据传递闭包的 λ-截矩阵得到聚类结果如图18(a) 所示. 同时, 画出动态聚类图如图18(b) 所示.



(a) λ-截矩阵所得的聚类结果　　(b) 动态聚类图

图 18: 基于 T_P-等价关系的聚类结果

计算出每个点的轮廓值, 以及画出相关性度量的热力图如下图所示.



(a) 轮廓值　　(b) 热力图

图 19: 基于 T_P-等价关系的聚类结果的评价

4.8 聚类效果的分析与对比

4.8.1 基于五种核函数的聚类结果对比

通过聚类结果图可以观察到, 利用 Gauss 核函数、Laplace 核函数、二次有理核函数及逆多元二次核函数得到的聚类结果是相同的, 计算得到的平均轮廓系数均为 0.3805; 而通过线性核函数的聚类结果, 虽然与上述四个结果中各类的元素数目相同, 但是具体元素存在差异, 且其平均轮廓系数仅有 0.0517, 效果相对较差且与上述四种方法差异明显. 此外, 通过观察热力图也可以发现, 线性核函数聚类得到的热力图中, 各类之间的相关性并不显著, 即深颜色的斑块并不明显, 且各板块颜色差异辨识度较小.

表 2: 基于五种核函数的聚类结果对比

核函数	等价关系类型	平均轮廓系数
线性核函数		0.0517
Gauss 核函数		0.3805
Laplace 核函数	$T_{\cos}$-等价关系	0.3805
二次有理核函数		0.3805
逆多元二次核函数		0.0517

综上所述, 在本次应用实践中, 基于核函数的聚类中, 采用线性核函数得到的结果最差, 其余四种方法无明显差异.

4.8.2 基于三种等价关系的聚类结果对比

在这一部分的讨论中, 利用 $T_{\cos}$-等价关系的五个聚类里不考虑基于线性核函数的聚类.

表 3: 基于三种等价关系的聚类结果对比

等价关系类型	平均轮廓系数
$T_{\cos}$-等价关系	0.3805
T_L-等价关系	0.1023
T_P-等价关系	0.0157

首先, 观察聚类结果图, 三种聚类方式得到的类中元素个数存在差异, 其中, 采用 T_L-等价关系进行聚类的结构与其余两种的差异较为明显. 而通过考察平均轮廓系数, 可

以发现, 由 $T_{\cos}$-等价关系、T_L-等价关系到 T_P-等价关系, 数值呈逐渐递减的趋势, 而相关性热力图呈现的对比度变化趋势与之相同, 即逐对比度逐渐变小. 根据前述定理, 这三种等价关系是逐渐增强的, 故而可以猜测是因为本次实践中使用的数据集较小, 过强的等价关系导致了过拟合, 故而轮廓系数减小. 若使用大样本数据, 几种等价关系的强弱应该会通过轮廓系数值体现出来.

5 总结

本次研究的主题是模糊等价关系与核函数之间的关系及其在动态模糊聚类中的应用.

在第一部分中, 我们介绍了一些基础知识: 核函数的定义、封闭性; t 模的定义及由此导出的蕴含及双蕴含关系; 在 t 模基础上构建的 T-等价关系的定义及 $T_{\cos}$、T_L、T_P、T_M-等价关系之间的强弱关系.

在第二部分中, 我们研究了核函数及 T-等价关系之间的关联, 指出了以下几个结论: 核函数是 $T_{\cos}$-等价关系, T_M-等价关系是核函数, 并给出了用 T_L、T_P-等价关系生成核函数的方法, 形成了完整的理论体系. 在 2.4 中, 为方便后续聚类中使用, 针对 T-等价关系的生成方法进行了综述.

在第三部分中, 对动态模糊聚类方法进行了概述, 对等价矩阵、λ-截矩阵、传递闭包及其存在性定理、计算方法进行了阐释, 随后给出了动态模糊聚类实施的具体步骤. 此外, 还给出了评价聚类结果的两个指标——轮廓值和相关性度量.

最后, 针对排球运动员的技术统计数据进行了聚类分析, 分别使用作为 $T_{\cos}$-等价关系的五类核函数及 T_L、T_P-等价关系进行聚类, 绘制聚类结果图、动态聚类图, 并结合轮廓数值图、相关性热力图对聚类效果进行了评价及分析.

结合本次实践得到的结果, 在较小的数据集下, 强等价关系的优点并没有得到较为明显的呈现, 因此, 下一步研究可从大数据样本入手, 设计针对大样本的算法, 并考察等价关系强弱在聚类效果上的体现.

总之, 在本次的研究中, 我们明确了核函数与 T-等价关系之间的关联, 由此拓宽了 T-等价关系的构建方法. 由此, 我们将有多种方法实现动态模糊聚类, 进而可以结合具体的实现效果, 选取效果最好的方法, 最大限度地挖掘数据集的特征, 为后续研究做准备.

参考文献

[1] Moser, Bernhard. *On the T-transitivity of kernels*. Fuzzy Sets and Systems, 2006, 157:1787-1796.

[2] Moser, Bernhard. *On Representing and Generating Kernels by Fuzzy Equivalence Relations*. Journal of Machine Learning Research, 2006, 12:2603-2620.

[3] Carl H. FitzGerald, Charles A. Micchelli, Allan Pinkus. *Functions that preserve families of positive semidefinite matrices*. Linear Algebra and its Applications, 1995, 221:83-102.

[4] Ling, C. H.. *Representation of Associative Functions*. Publicationes Mathematicae, 1995:12, 189-212.

[5] Erich Peter Klement, Radko Mesiar, Endre Pap. *Triangular Norms*. Kluwer Academic Publishers, Dordrecht, 2000.

[6] Murat Yetis, Gedizlioglu Ergun. *A new approach for fuzzy traffic signal control*. 1995.

[7] Trillas E., Valverde L. *An Inquiry into Indistinguishability Operators*. Aspects of Vagueness. 1984:39, 231-256.

[8] Rudolf Kruse, Jörg Gebhardt, Frank Klawonn. *Fuzzy-Systeme*. 1993.

[9] Rudolf Kruse, Joan E. Gebhardt, F. Klowon. *Foundations of Fuzzy Systems*. John Wiley & Sons, Inc., 1994.

[10] Bezdek J. C.. *A convergence theorem for the fuzzy ISODATA clustering algorithms*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1980:2, 1-8.

附录

A 所用软件

论文的排版使用了基于 Visual Studio Code 的 LaTeX, 模板为 ElegantPaper, 关于 ElegantLaTeX 系列的模板可见ElegantLaTeX 系列模板.

图像的绘制和代码的编写使用了 MathWorks Matlab.

B 代码

B.1 绘制核函数的图像

```
    xspan = -1 : 0.05 : 1;
    yspan = -1 : 0.05 : 1;
    [x, y] = meshgrid(xspan, yspan);

```

```
    figure();
    colormap Winter;

    k1 = x .* y;
    subplot(1, 2, 1);
    surf(x, y, k1);
    title('线性核函数的图像');
    axis equal;

    k2 = exp(-1/2 .* abs(x - y).^2);
    subplot(1, 2, 2);
    surf(x, y, k2);
    title('Gauss核函数的图像');
    axis equal;

    figure();
    colormap Winter;

    k3 = exp(-1 .* abs(x - y));
    subplot(2, 2, 1);
    surf(x, y, k3);
    title('Laplace核函数的图像');
    axis equal;

    k4 = 1 ./(abs(x-y).^2 + 1);
    subplot(2, 2, 2);
    surf(x, y, k4);
    title('二次有理核函数的图像');
    axis equal;

    k5 = sqrt(abs(x-y).^2 + 1);
    subplot(2, 2, 3);
    surf(x, y, k5);
    title('多元二次核函数的图像');
    axis equal;

    k6 = 1 ./ sqrt(abs(x-y).^2 + 1);
```

```
42      subplot(2, 2, 4);
43      surf(x, y, k6);
44      title('逆多元二次核函数的图像');
45      axis equal;
```

B.2 绘制 t 模的图像

```
1       xspan = 0 : 0.05 : 1;
2       yspan = 0 : 0.05 : 1;
3       [x, y] = meshgrid(xspan, yspan);
4       colormap Winter;
5
6       Tcos = max(x.*y - sqrt(1-x.^2).*sqrt(1-y.^2), 0);
7       subplot(2, 2, 1);
8       surf(x, y, Tcos);
9       title('Tcos的图像');
10      axis equal;
11
12      TL = max(x + y - 1, 0);
13      subplot(2, 2, 2);
14      surf(x, y, TL);
15      title('TL的图像');
16      axis equal;
17
18      TP = x.*y;
19      subplot(2, 2, 3);
20      surf(x, y, TP);
21      title('TP的图像');
22      axis equal;
23
24      TM = min(x, y);
25      subplot(2, 2, 4);
26      surf(x, y, TM);
27      title('TM的图像');
28      axis equal;
```

B.3 动态模糊聚类与结果的评价

```
%% 初始化

clc;
clear;
close all;

%% T-等价关系

% E = @(x, y) sum(x.*y) / (norm(x)*norm(y)); % 线性核函数, ...
    Tcos-等价关系
% sigma = 1;
% E = @(x, y) exp(-1/(2 * sigma^2) * norm(x - y)^2); % ...
    Gauss核函数, Tcos-等价关系
% E = @(x, y) exp(-1/sigma * norm(x-y)); % Laplace核函数, ...
    Tcos-等价关系
c = 1;
% E = @(x, y) 1 - norm(x - y)^2/(norm(x-y)^2 + c); % ...
    二次有理核函数, Tcos-等价关系
% E = @(x, y) 1/sqrt(norm(x - y)^2 + c^2); % 逆多元二次核函数, ...
    Tcos-等价关系

mu = @(x) sum(x)/16;
% E = @(x, y) min(mu(x)/mu(y), mu(y)/mu(x)); % TP-等价关系
E = @(x, y) 1 - abs(mu(x) - mu(y)); % TL-等价关系

%% 数据预处理

data = xlsread('/Volleyball.csv');
x = data(:, 2 : 16);
x = (x - min(x).*ones(size(x)))./(max(x).*ones(size(x)) - ...
    min(x).*ones(size(x))); % 用极差进行正规化
```

```
26
27      %% 用T-等价关系求相似矩阵
28
29      n = size(x, 1); % 数据量
30      r = zeros(n, n);
31      for i = 1 : n
32          for j = 1 : n
33              r(i, j) = E(x(i, :), x(j, :));
34          end
35      end
36
37      %% 求传递闭包
38
39      for k = 1 : ceil(log2(n)) % 最多需要计算n次
40          for i = 1 : n
41              for j = 1 : n
42                  r(i, j) = max(min(r(i, :), r(:, j)'));
43              end
44          end
45      end
46
47      %% 求lambda-截矩阵
48
49      c = 5; % 分类数
50      l = unique(r(:)); % lambda的取值范围
51      lambda = l(c);
52      R = zeros(n, n);
53      for i = 1 : n
54          for j = 1 : n
55              if r(i, j) >= lambda
56                  R(i, j) = 1;
57              else
58                  R(i, j) = 0;
59              end
60          end
61      end
62
```

```
63      %% 利用以上结果画图
64
65      figure();
66      G = graph(R, 'omitselfloops'); % 生成图
67      plot(G);%蓝色的图就是说
68
69      %% 动态聚类图
70
71      figure();
72      tree = linkage(r);
73      m = size(l);
74      for i = 1 : n - 1 % 令tree的第三列等于1 - lambda
75          if n - i > m
76              tree(i, 3) = 0;
77          else
78              tree(i, 3) = 1 - l(n - i);
79          end
80      end
81      dendrogram(tree);
82      result = cluster(tree, 'maxclust', c)';
83
84      %% 聚类结果的评价
85
86      figure();
87      silhouette(x, result) % 画出评价图, ...
            evaluation越接近1表示该簇的聚类效果越好
88
89      [col1] = find(result == 1, 30);
90      [col2] = find(result == 2, 30);
91      [col3] = find(result == 3, 30);
92      [col4] = find(result == 4, 30);
93      [col5] = find(result == 5, 30);
94
95      X = x';
96      col = {col1, col2, col3, col4, col5};
97      temp = zeros(size(X));
98      temp = [X(:,col{1,1}), X(:,col{1,2}), X(:,col{1,3}), ...
```

```
            X(:,col{1,4}), X(:,col{1,5})];
99
100     figure();
101     rho = corr(temp);
102     heat = heatmap(rho);
```