



可拓集、可拓策略生成和 可拓数据挖掘

汤 龙 博士

广东工业大学可拓学与创新方法研究所 副教授
中国人工智能学会可拓学专业委员会
常务委员、副秘书长

20210526

摘要

- 可拓集作为可拓学的三大基础理论之一，旨在描述事物的可变性，并通过建立实域上的关联函数，来定量地刻画对象具有或不具有某种性质的程度。本报告分为两部分内容：第一，从可拓集的分类思想出发，介绍可拓集的概念、定义、应用案例，以及关联函数的构建思想和方法；第二，简要介绍可拓集的两个应用方向——可拓策略生成和可拓数据挖掘，以及近年来的相关研究成果。

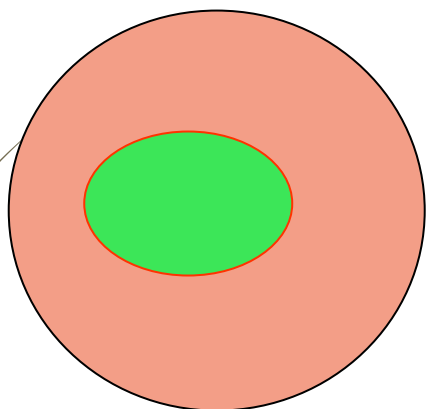
目录

- 1. 可拓集
- 2. 关联函数
- 3. 应用方向之一——可拓策略生成
- 4. 应用方向之二——可拓数据挖掘

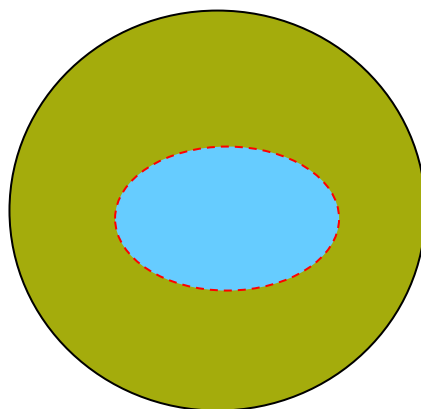
1. 可拓集

4

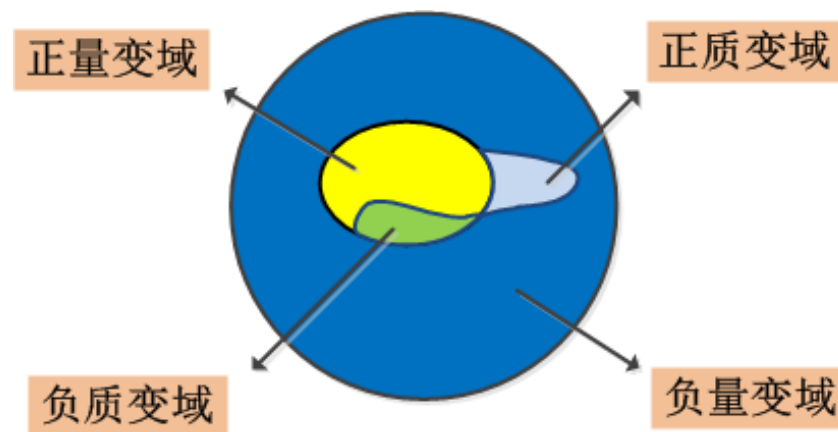
三类集合的比较：



经典集



模糊集



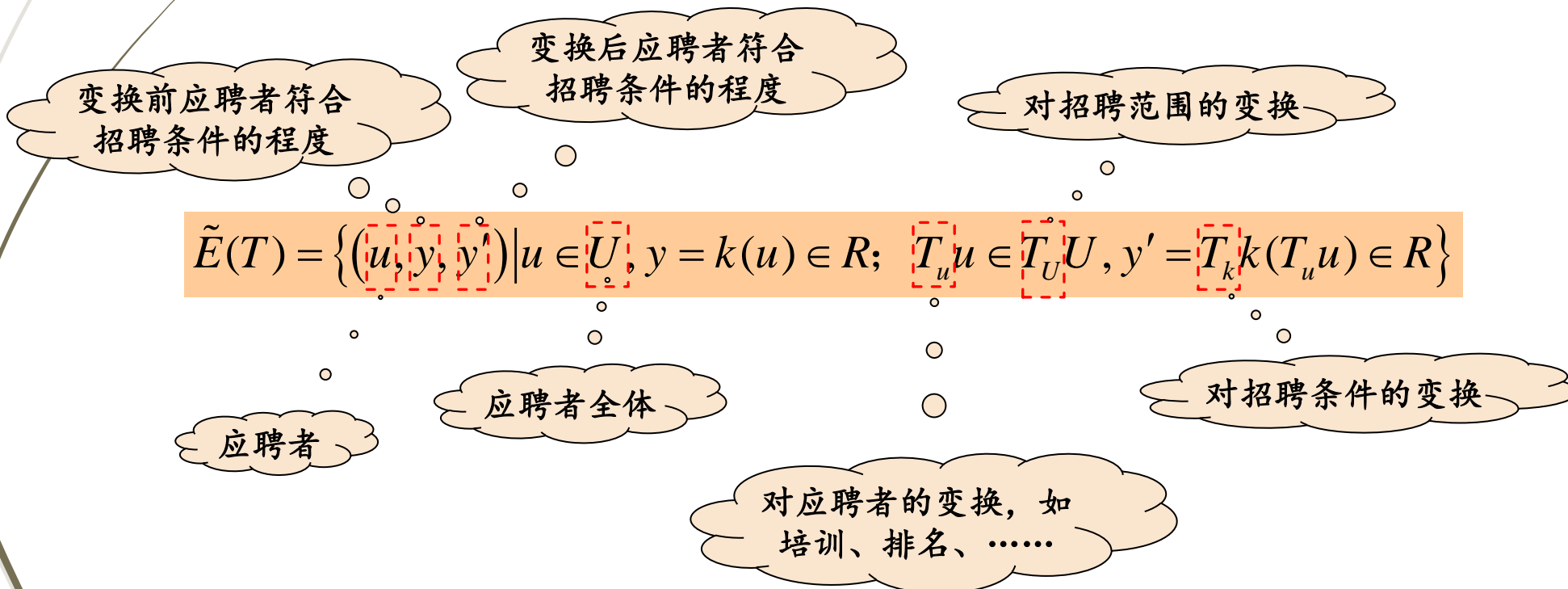
可拓集

1. 可拓集

5

以招聘人才为例

- ➔ 以某企业招聘人才为例，设论域 U 为应聘者的全体， $u \in U$ 为任一应聘者， $y=k(u)$ 表示应聘者 u 符合招聘条件的程度，则论域 U 的可拓集为



1. 可拓集

6

(1) 在不实施变换 T ($T=e$) 时,

➡ **正域**为

$$E_+ = \{ (u, y) \mid u \in U, y = k(u) > 0 \}$$

表示参选者中所有符合招聘条件的人员的全体。

➡ **负域**为

$$E_- = \{ (u, y) \mid u \in U, y = k(u) < 0 \}$$

表示参选者中所有不满足招聘条件的人员的全体。

➡ **零界**为

$$E_0 = \{ (u, y) \mid u \in U, y = k(u) = 0 \}$$

个人认为：零界的实际意义与招聘条件有关，若招聘条件为拥有某一证书，零界包括已经考到某一证书，但证书还未颁发下来的人员；若招聘条件为**专项技能考试排名的前若干名**，则包括排名与符合条件的末位名次相同的人员。

1. 可拓集

7

(2) 对元素做变换 Tu ——对所有应聘者进行专项技能培训一个月，则：

- **正质变域**表示原来专项技能水平排名靠后但培训后专项技能水平排名提前，由不符合条件者变为符合条件者的应聘者的全体。由于这些人变为符合条件，但名额有限，会使原来部分符合条件的应聘者排名退后而被淘汰。
- **负质变域**表示原来符合条件，但后来又被淘汰的应聘者的全体。
- **正量变域**表示原来符合条件，经变换 Tu 后仍然符合条件的应聘者的全体。
- **负量变域**表示原来不符合条件，经变换 Tu 后仍然不符合条件的应聘者的全体。

1. 可拓集

8

(3) 对关联准则做变换 T_k ——将招聘条件由原来的专项技能水平排名前10, 变换为排名前 $n(n>10$ 或 $n<10)$, 则:

- **正质变域**表示原来不符合条件, 但变换招聘条件后变为符合条件的应聘者的全体 (条件从 $n=10$ 变换为 $n>10$ 时增加的符合新招聘条件的人员);
- **负质变域**表示原来符合条件, 但变换招聘条件后变为不符合条件的应聘者的全体 (条件从 $n=10$ 变换为 $n<10$ 时减少的不符合新招聘条件的人员);
- **正量变域**表示原来符合条件, 经变换后仍然符合条件的应聘者的全体 (既符合原招聘条件又符合新招聘条件的人员);
- **负量变域**表示原来不符合条件, 经变换后仍然不符合条件的应聘者的全体 (既不符合原招聘条件又不符合新招聘条件的人员) 。

1. 可拓集

9

(4) 若变换 TU 为对论域 U 的变换, 则可拓集为

$$\tilde{E}(T_U) = \{ (u, y, y') \mid u \in U, y = k(u) \in R; u \in T_U U, y' = k'(u) \in R \}$$

- 设 TU 为扩大论域区域, 如原来的论域只是广州市, 现在变为广东省, 招聘条件不变, 则:
- **正质变域**表示非广州市的应聘者, 经变换后变为符合条件的应聘者的全体;
- **负质变域**表示广州市内原来符合条件的参选者, 由于外市人员的加入而变为不符合条件的参选者的全体;
- **正量变域**表示原来符合条件, 经变换后仍然符合条件的应聘者的全体;
- **负量变域**表示原来不符合条件, 经变换后仍然不符合条件的应聘者的全体。

1. 可拓集

10

可拓集的定义

$$\tilde{E}(T) = \{(u, y, y') \mid u \in U, y = k(u) \in R; \overset{\circ}{T}_u u \in \overset{\circ}{T}_U U, y' = \overset{\circ}{T}_k k(T_u u) \in R\}$$

对元素的
变换

对论域的
变换

对关联准则
的变换

\Downarrow

\Downarrow

\Downarrow

u'

U'

k'

变换后
的元素

变换后
的论域

变换后的
关联准则

$$\tilde{E}(T) = \{(u, y, y') \mid u \in U, y = k(u) \in R, u' \in U', y' = k'(u') \in R\}$$

1. 可拓集

用可拓集表示质变、量变和临界状态

► 当 $T \neq e$ 时,

$$\dot{E}_+(T) = \{(u, y, y') | u \in U, T_u u \in T_U U, y \leq 0, y' > 0\}$$

称为正质变域;

$$\dot{E}_-(T) = \{(u, y, y') | u \in U, T_u u \in T_U U, y \geq 0, y' < 0\}$$

称为负质变域;

$$E_+(T) = \{(u, y, y') | u \in U, T_u u \in T_U U, y > 0, y' > 0\}$$

称为正量变域;

$$E_-(T) = \{(u, y, y') | u \in U, T_u u \in T_U U, y < 0, y' < 0\}$$

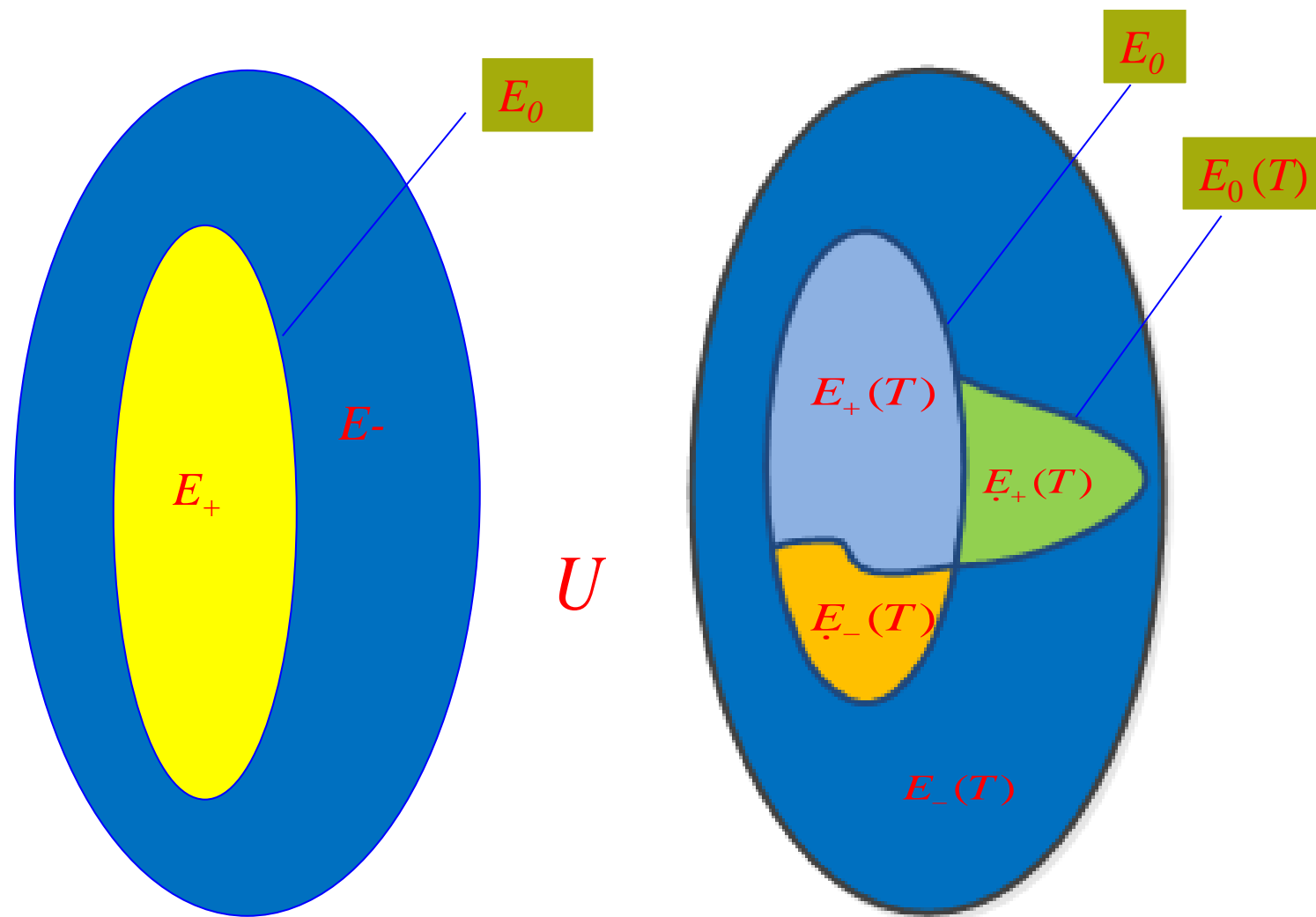
称为负量变域。

$$E_0(T) = \{(u, y, y') | u \in U, T_u u \in T_U U, y' = 0\}$$

称为拓界。

1. 可拓集

12



当 $T=e$ 时对论域 U 的划分

关于元素变换的可拓集对论域的划分

2. 关联函数

13

关联函数的意义

- **关联函数** $k(x)$ ——刻画论域中的元素具有某种性质的程度。
- 如基元可拓集中，对象在某评价特征上的“矛盾的程度”或“符合要求的程度”
 - $k(x)$ 一定是对于某个**评价特征**（如：成本，材料，形状）而言的， x 表示研究对象关于这一评价特征的量值
 - $k(x)>0$ 表示研究对象关于某评价特征符合要求，即属于正域
 - $k(x)<0$ 表示研究对象关于某评价特征不符合要求，即属于负域
 - $k(x)=0$ 表示研究对象关于某评价特征处于临界状态

2. 关联函数

14

关联函数的类型

- 连续型关联函数（如基元可拓集中，评价特征的量值连续，如：重量，高度，温度，电流）
- 简单关联函数
- 初等关联函数（基本概念、意义、用法）
- 离散型关联函数（如基元可拓集中，评价特征的量值离散，如：创新程度，耗电等级）

2. 关联函数

15

简单关联函数

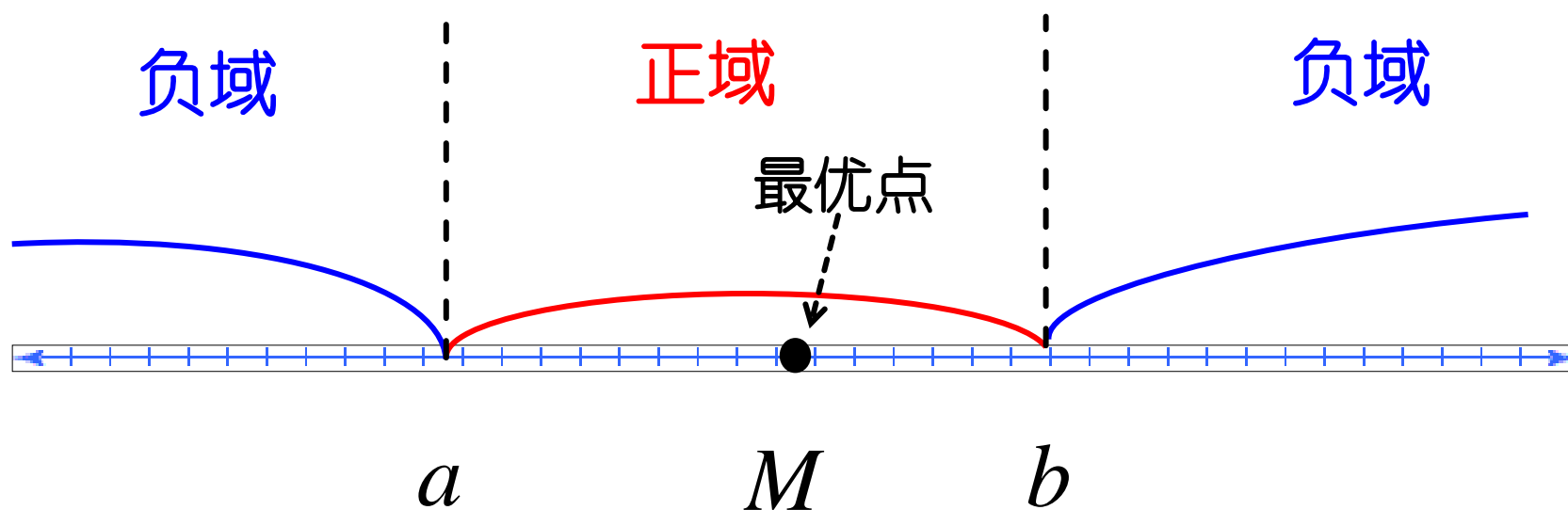
- **正域为有限区间**
- **背景实例：**
 - 铁轨转弯处对火车运行速度的要求
 - 工厂加工工件的合格范围

2. 关联函数

16

简单关联函数

- 正域为有限区间
- 正域和负域的划分：



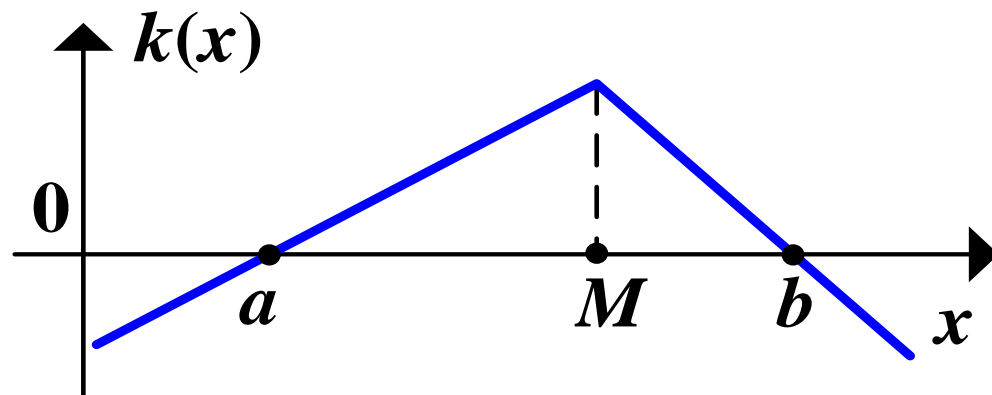
2. 关联函数

17

简单关联函数

➤ 正域为有限区间

➤ 函数示意图：



➤ 在区间 $\langle a, b \rangle$ 内部, $k(x) > 0$;

➤ 在区间 $\langle a, b \rangle$ 外部, $k(x) < 0$;

➤ 在区间 $\langle a, b \rangle$ 端点, $k(x) = 0$;

➤ 在最优点 M 处, $k(x)$ 取得最大值;

2. 关联函数

18

初等关联函数

➡ 背景实例：

➡ 孔子的“中庸之道”——儒家的道德标准，待人接物不偏不倚，调和折中

不及则倾，中庸则正，过之则覆

➡ 额定电压；额定电流

2. 关联函数

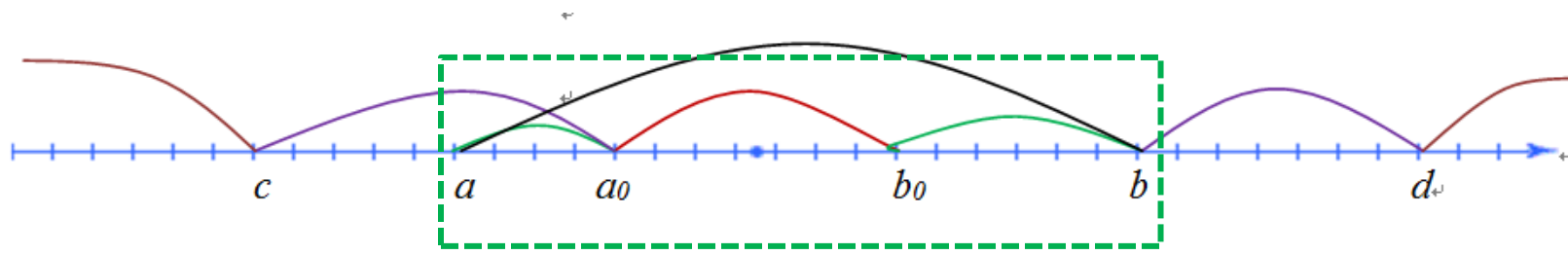
19

初等关联函数

初等关联函数对论域的划分:

$$\text{论域 } R \begin{cases} \text{正域: } \langle a, b \rangle \begin{cases} \text{标准正域 } \langle a_0, b_0 \rangle \\ \text{过渡正域 } \langle a, a_0 \rangle \cup \langle b_0, b \rangle \\ \text{最优点 } x_0 \in X_0 \end{cases} \\ \text{零界 } a, b \\ \text{负域: } R - \langle a, b \rangle \begin{cases} \text{标准负域 } \langle -\infty, a \rangle \cup \langle b, +\infty \rangle \\ \text{过渡负域 } \langle c, a \rangle \cup \langle b, d \rangle \end{cases} \end{cases}$$

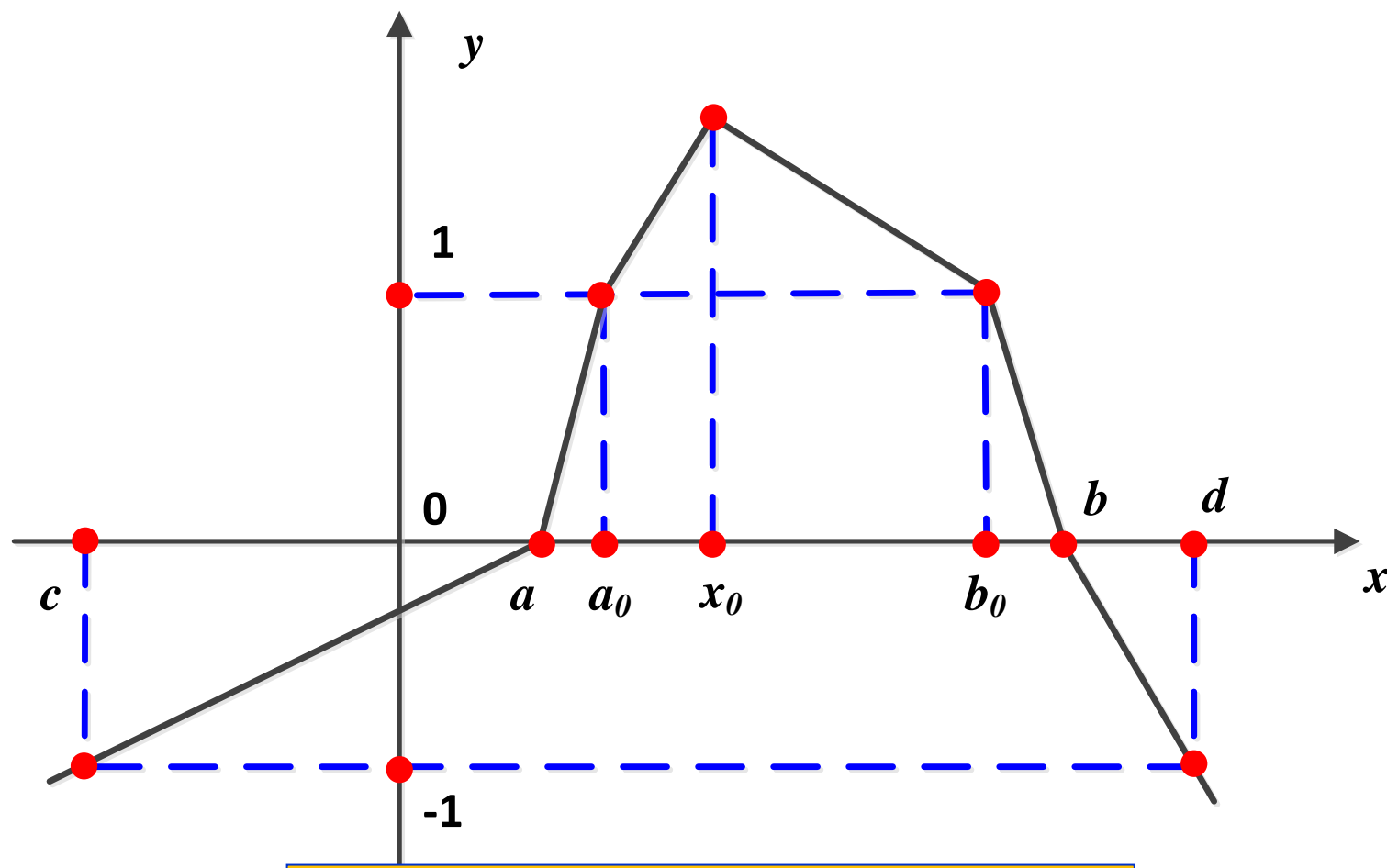
$$X_0 = \langle a_0, b_0 \rangle, X = \langle a, b \rangle, \hat{X} = \langle c, d \rangle$$



2. 关联函数

20

初等关联函数



三区间套初等关联函数 $y=k(x)$ 示意图

2. 关联函数

21

初等关联函数

标准正域, $k(x) \geq 1$; 过渡正域, $0 \leq k(x) \leq 1$;

过渡负域, $-1 \leq k(x) \leq 0$; 标准负域, $k(x) \leq -1$;

$k(x)$ 在最优点 M 取得最大值;

$$k(x) = \begin{cases} \frac{\rho(x, M, X)}{D(x, M, X_0, X)}, & D(x, M, X_0, X) \neq 0, x \in X \\ -\rho(x, M, X_0) + 1, & D(x, M, X_0, X) = 0, x \in X_0 \\ 0, & D(x, M, X_0, X) = 0, x \notin X_0, x \in X \\ \frac{\rho(x, M, X)}{D(x, M, X, \hat{X})}, & D(x, M, X, \hat{X}) \neq 0, x \in \mathfrak{R} - X \\ -\rho(x, M, \hat{X}) + 1, & D(x, M, X, \hat{X}) = 0, x \in \mathfrak{R} - X \end{cases}$$

2. 关联函数

22

初等关联函数

实例：某零件的理想尺寸为10mm

尺寸在9.95mm到 10.05mm之间的为优品

尺寸在9.9mm到 10.1mm之间的为合格品

尺寸在9.8mm到9.9mm之间以及10.1mm到10.2m之间的为可返工品

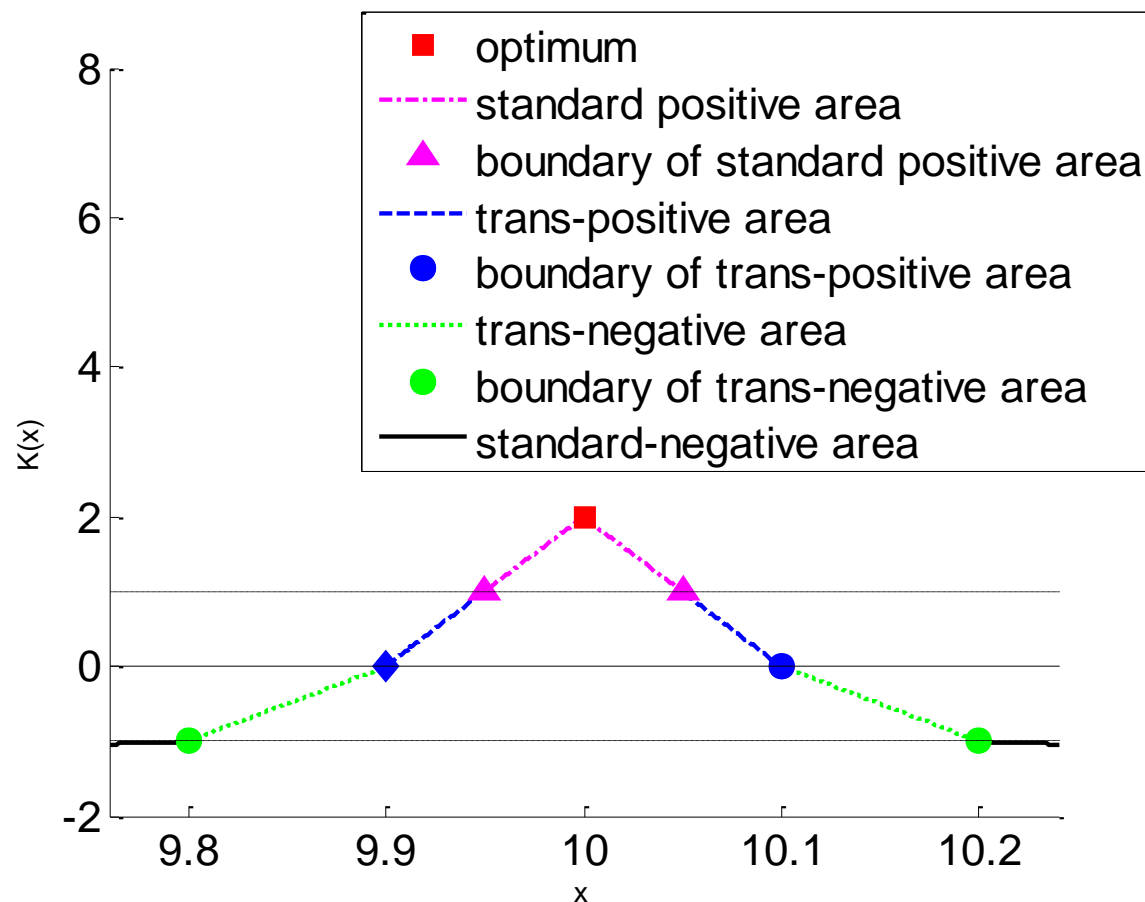
试建立产品关于“尺寸”的初等关联函数所对应的三区间套。

2. 关联函数

23

初等关联函数

$X_0 = \langle 9.95, 10.05 \rangle$, $X = \langle 9.9, 10.1 \rangle$, $\hat{X} = \langle 9.8, 10.02 \rangle$,
最优点 $M=10$, 初等关联函数 $k(x)$,

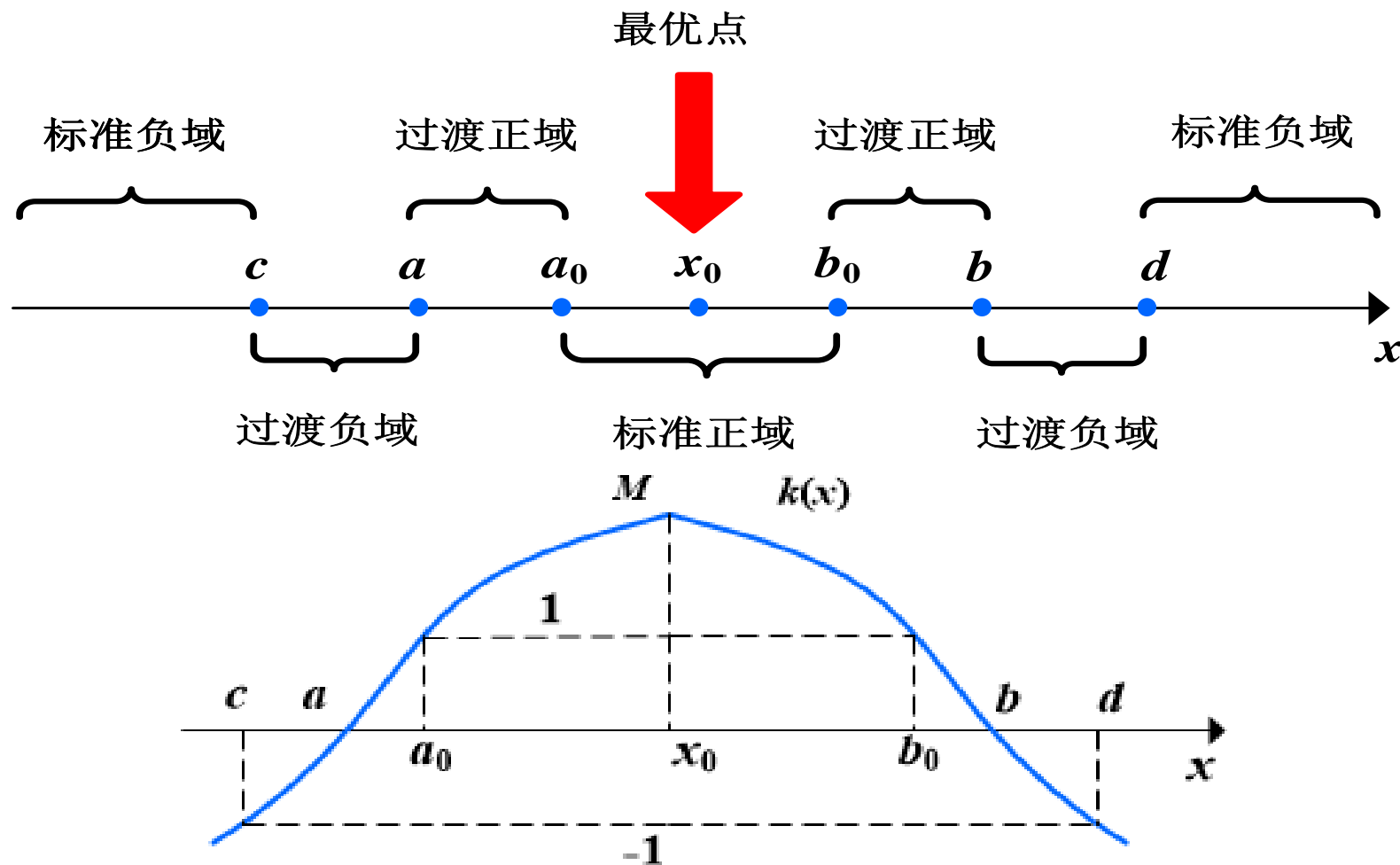


2. 关联函数

24

基于区间变换的多模式初等关联函数构建方法

- 基于可拓距的初等关联函数——函数模式单一。如何根据问题的特点构建具有相应模式的初等关联函数？（线性，非线性）

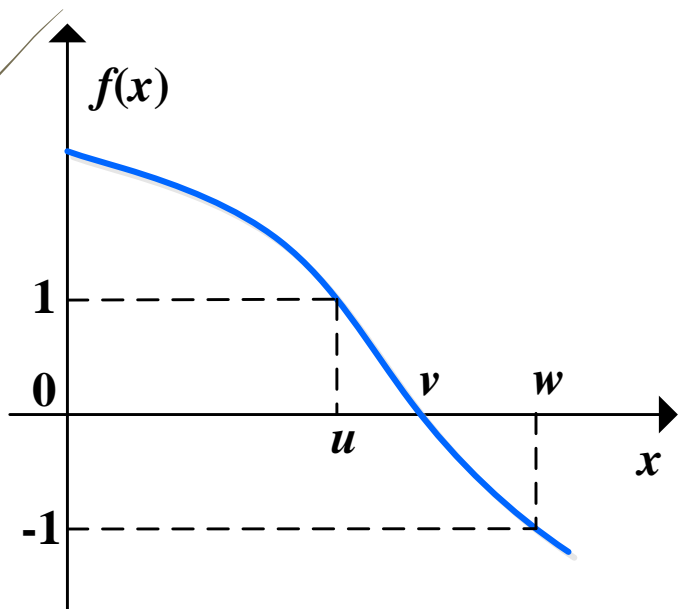


2. 关联函数

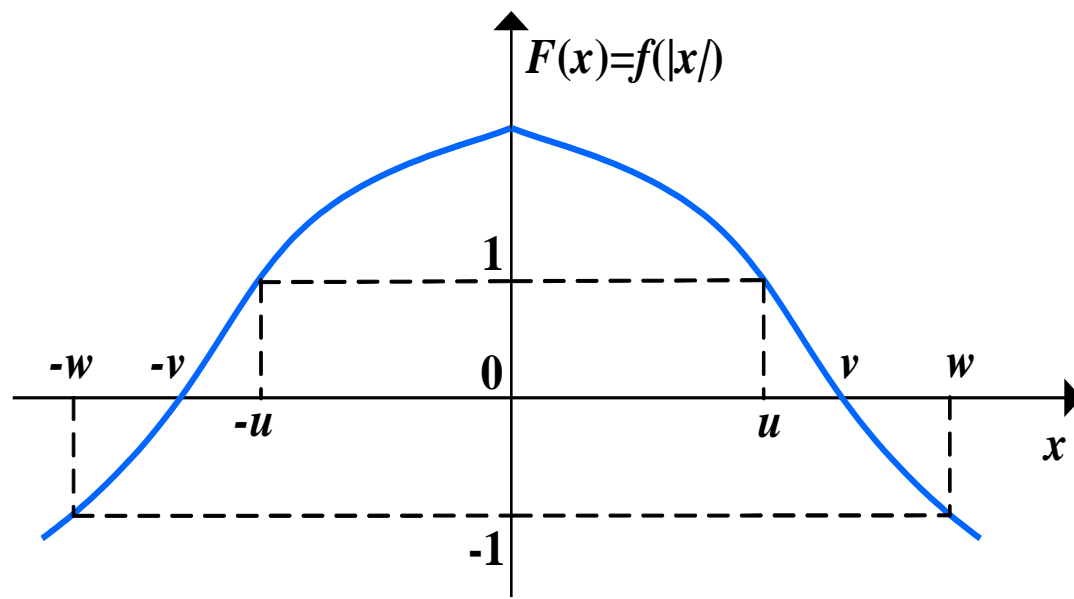
25

基于区间变换的多模式初等关联函数构建方法

- 基函数与标准函数



基函数



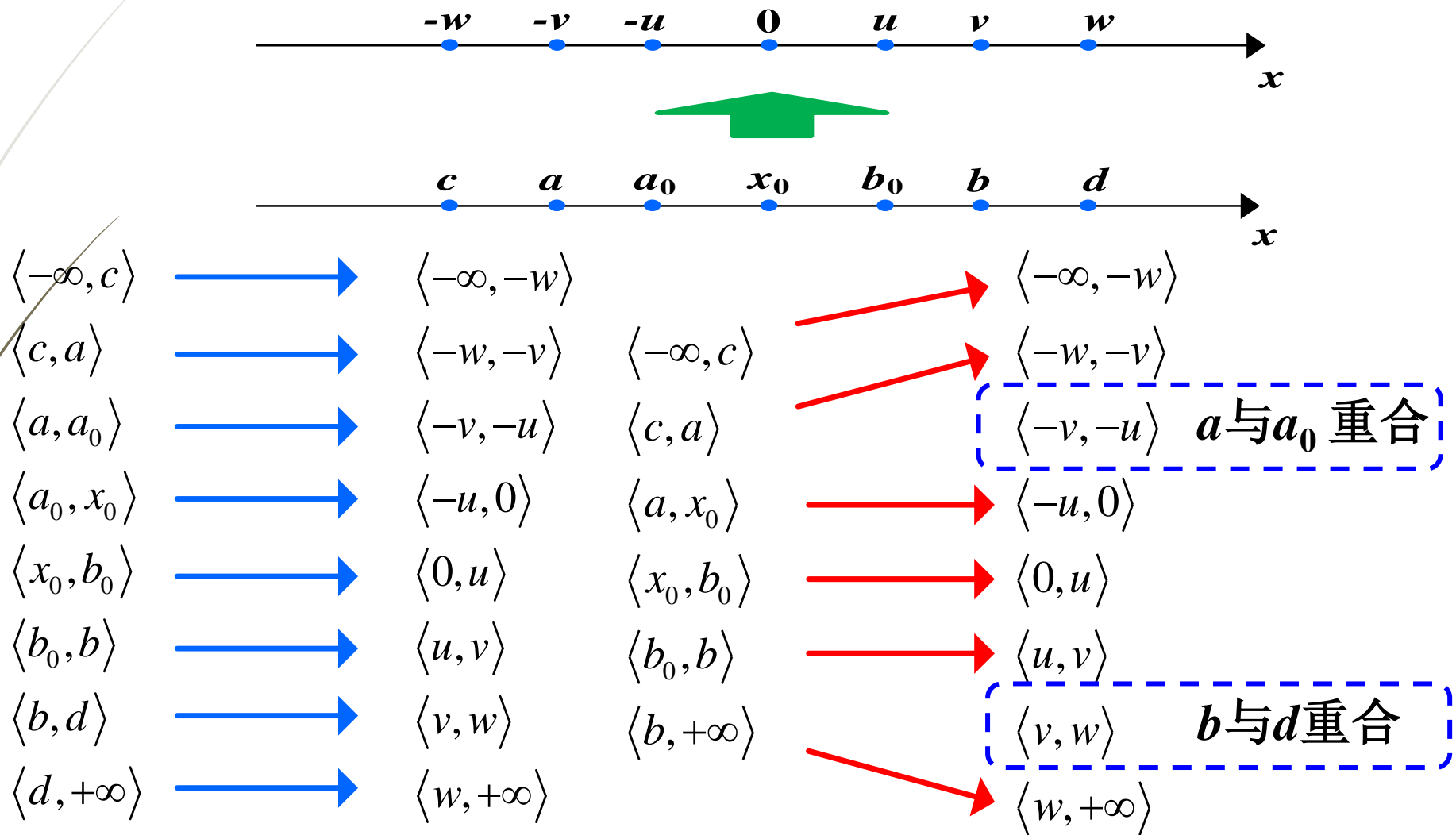
标准函数

2. 关联函数

26

基于区间变换的多模式初等关联函数构建方法

- 标准区间变换

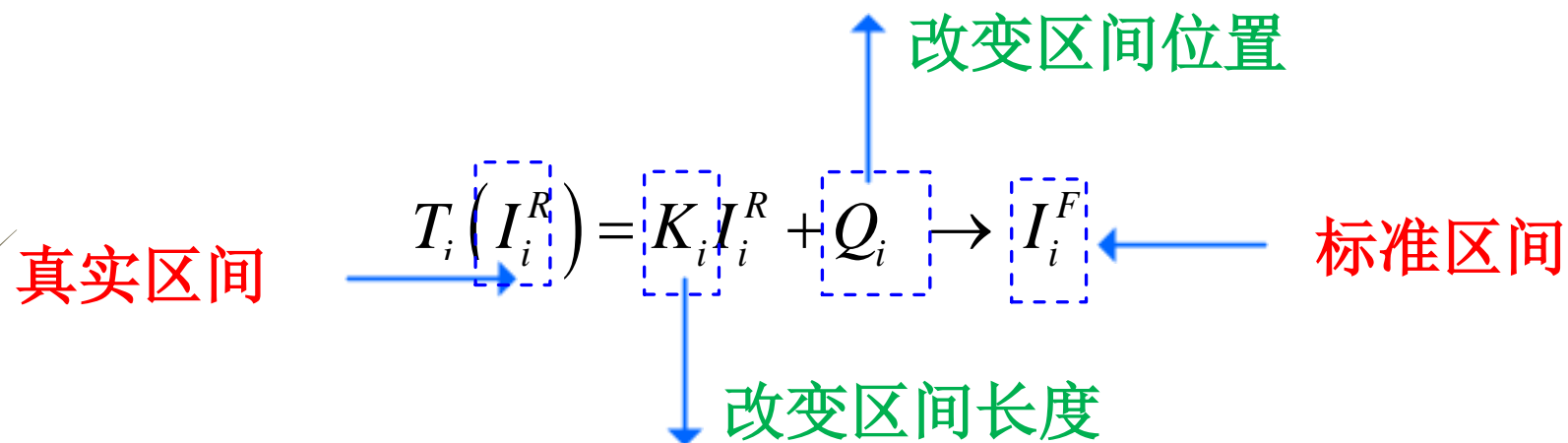


2. 关联函数

27

基于区间变换的多模式初等关联函数构建方法

- 标准区间变换



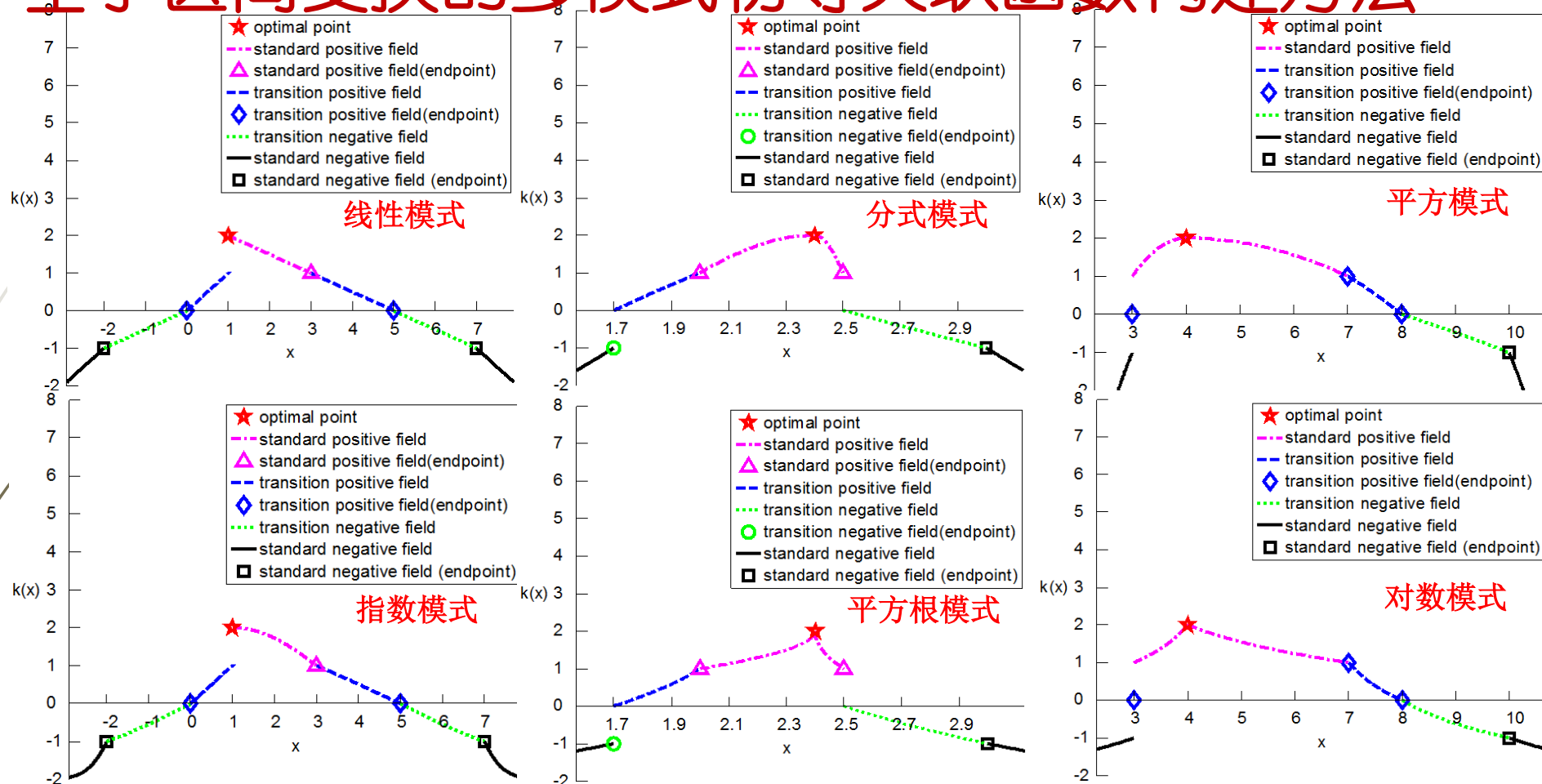
$$\begin{pmatrix} I_{il}^R & 1 \\ I_{ir}^R & 1 \end{pmatrix} \begin{pmatrix} K_i \\ Q_i \end{pmatrix} = \begin{pmatrix} I_{il}^F \\ I_{ir}^F \end{pmatrix} \Rightarrow \begin{pmatrix} K_i \\ Q_i \end{pmatrix} = \begin{pmatrix} I_{il}^R & 1 \\ I_{ir}^R & 1 \end{pmatrix}^{-1} \begin{pmatrix} I_{il}^F \\ I_{ir}^F \end{pmatrix}$$

$$k(x) = F(K_i x + Q_i) = f(|K_i x + Q_i|), x \in I_i^R$$

2. 关联函数

28

基于区间变换的多模式初等关联函数构建方法



Long Tang*, Chunyan Yang, A novel establishing method for 1-dimensional elementary dependent function by standard interval transformation. *Advances in Intelligent and Soft Computing*, 2016, 443, pp.563-575 (EI).

广东工业大学, 多模式初等关联函数计算软件V1.0, 2018SR029309, 原始取得, 全部权利, 2017.09.28.

3. 可拓集的应用——可拓策略生成

29

人们希望通过变换，把原来不符合要求的变为符合要求的，即构建正质变域，如何实现？？

解决矛盾问题的关键在于变换，但变换的方式和路径千千万万，**怎么办？？**

如何用最简单的方式表示这些**变换**？？

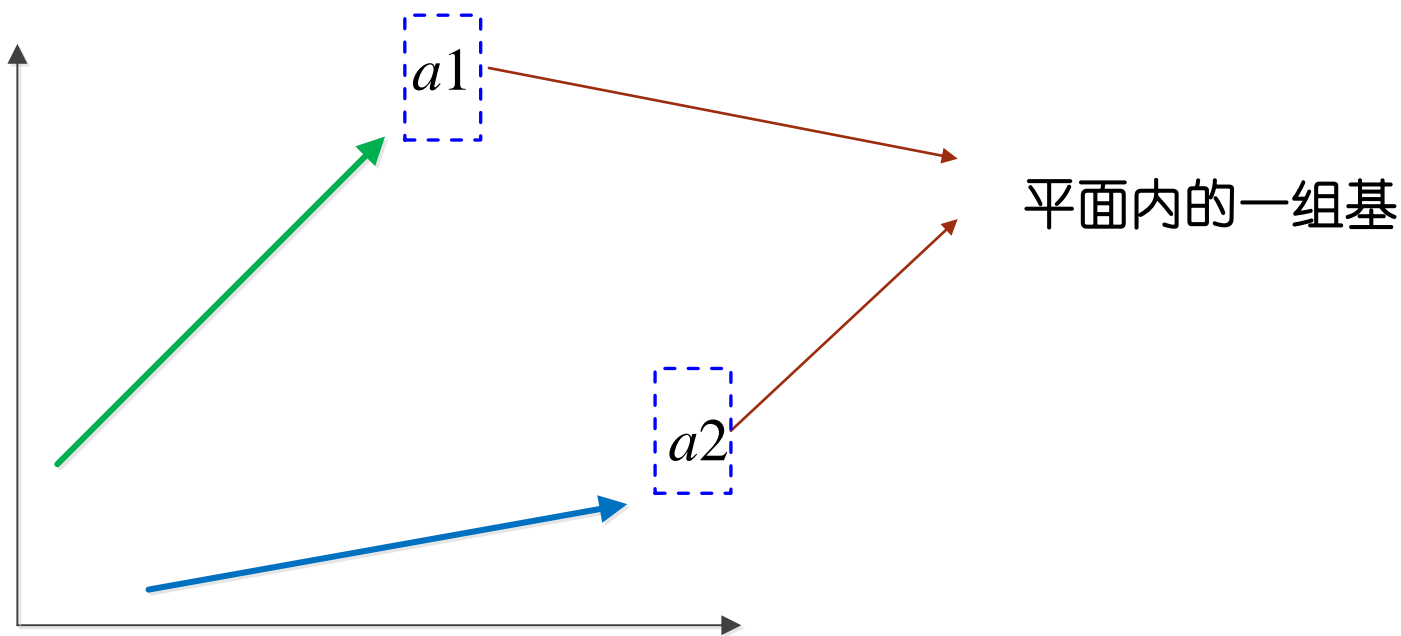
3. 可拓集的应用——可拓策略生成

30

举例：如何表达平面内的所有向量？

长度： $[0, +\infty)$

方向： $[0, 2\pi]$



$$c_1 a_1 + c_2 a_2, \quad c_1, c_2 \in R$$

3. 可拓集的应用——可拓策略生成

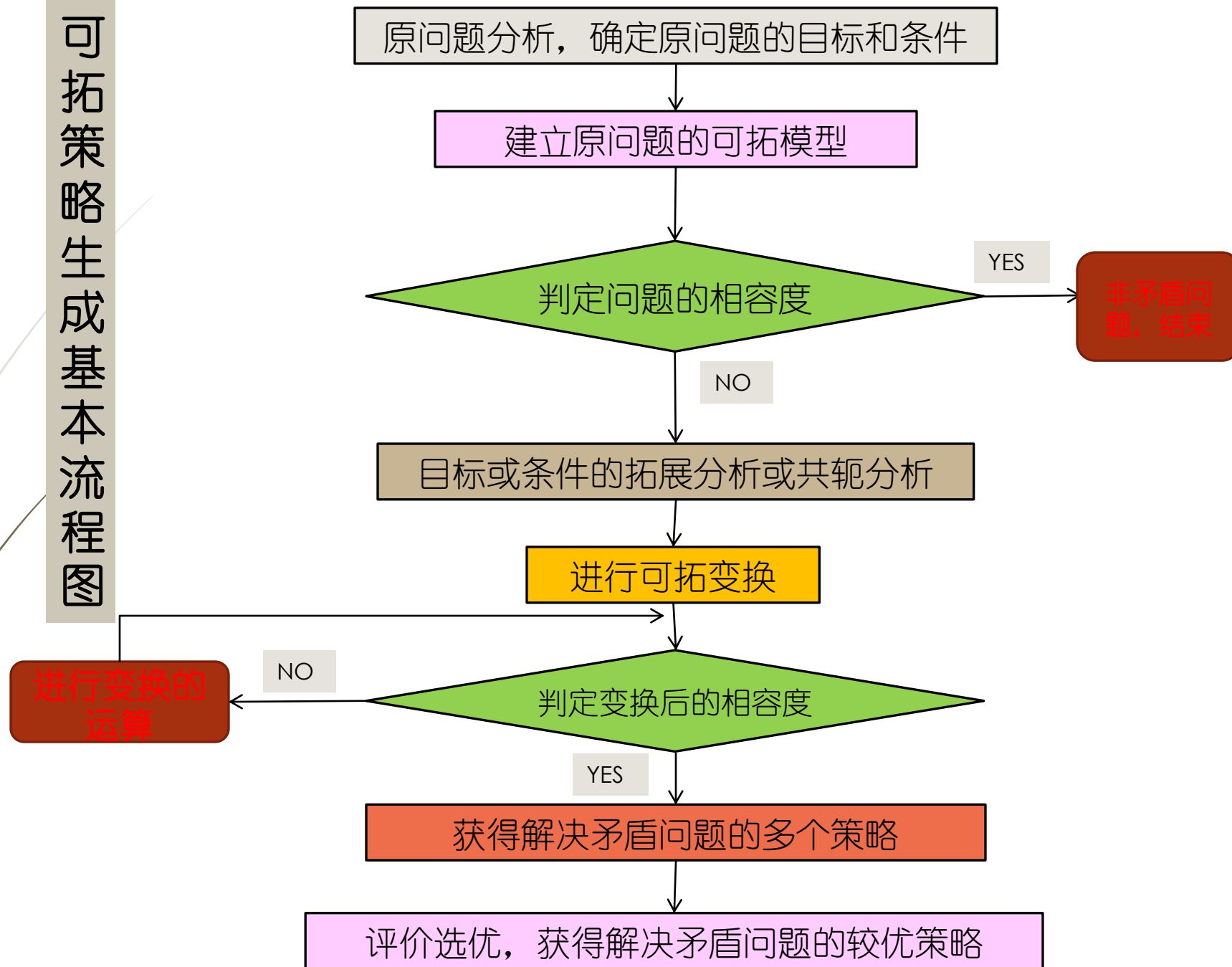
31

类比：如何找到变换空间中的一组“基”？

五种基本变换：置换、增删、扩缩、分解、复制

四种基本运算：积、与、或、逆

可拓策略生成基本流程图

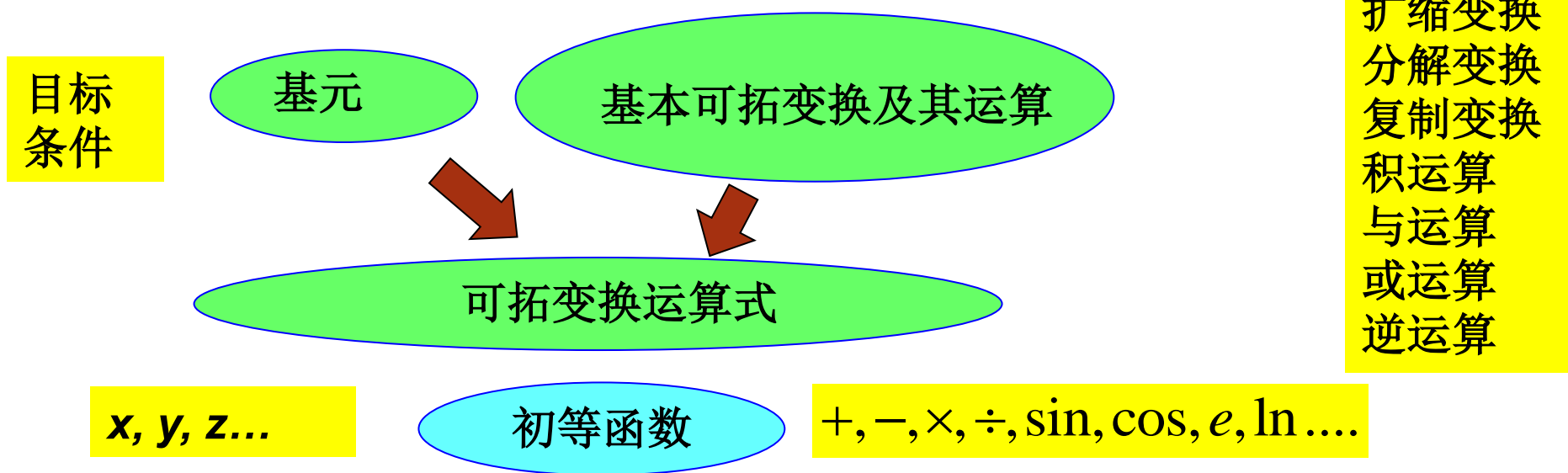


3. 可拓集的应用——可拓策略生成

33

基于GEP的可拓变换运算式自组织构建机制

- 如何构建可拓变换的运算式：现有根据问题目标和条件中属性（特征）的属性值（特征值）的相容程度来选择基本可拓变换的类型以及变换后的基元。
- 复杂不相容问题需要利用传导变换来求解时，主动可拓变换和变换结果之间往往会呈现出较强的“黑箱”特性。



- 随着大数据时代的到来和领域知识的不断完善，可拓策略生成过程中可拓变换运算式的数量面临组合爆炸。

3. 可拓集的应用——可拓策略生成

34

基于GEP的可拓变换运算式自组织构建机制

- 每一个染色体都代表一组可拓变换运算式；种群结构在个体选择机制和遗传算子的作用下朝好的方向进化，种群中个体的相容度和优度也随之提升，直至收敛。
- **染色体**：由代表不同变换对象可拓变换运算式的基因序列构成。
- **变换的对象**

设不相容问题核问题的可拓模型：

$$p_0 = g_0 * l_0$$

变换的对象

对条件 l_0 作相关分析

$$l_0 \Leftarrow \begin{cases} l_1 \\ l_2 \\ \dots \\ l_n \end{cases}$$

$$g_0 = \begin{bmatrix} Z_{g0}, & c_{0s1}, & c_{0s1}(Z_{g0}) \\ c_{0s2}, & c_{0s2}(Z_{g0}) \\ \dots & \dots \\ c_{0sn}, & c_{0sn}(Z_{g0}) \end{bmatrix}, \quad l_0 = \begin{bmatrix} Z_{l0}, & c_{0r1}, & c_{0r1}(Z_{l0}) \\ c_{0r2}, & c_{0r2}(Z_{l0}) \\ \dots & \dots \\ c_{0rn}, & c_{0rn}(Z_{l0}) \end{bmatrix}$$

得到条件相关树

叶基元

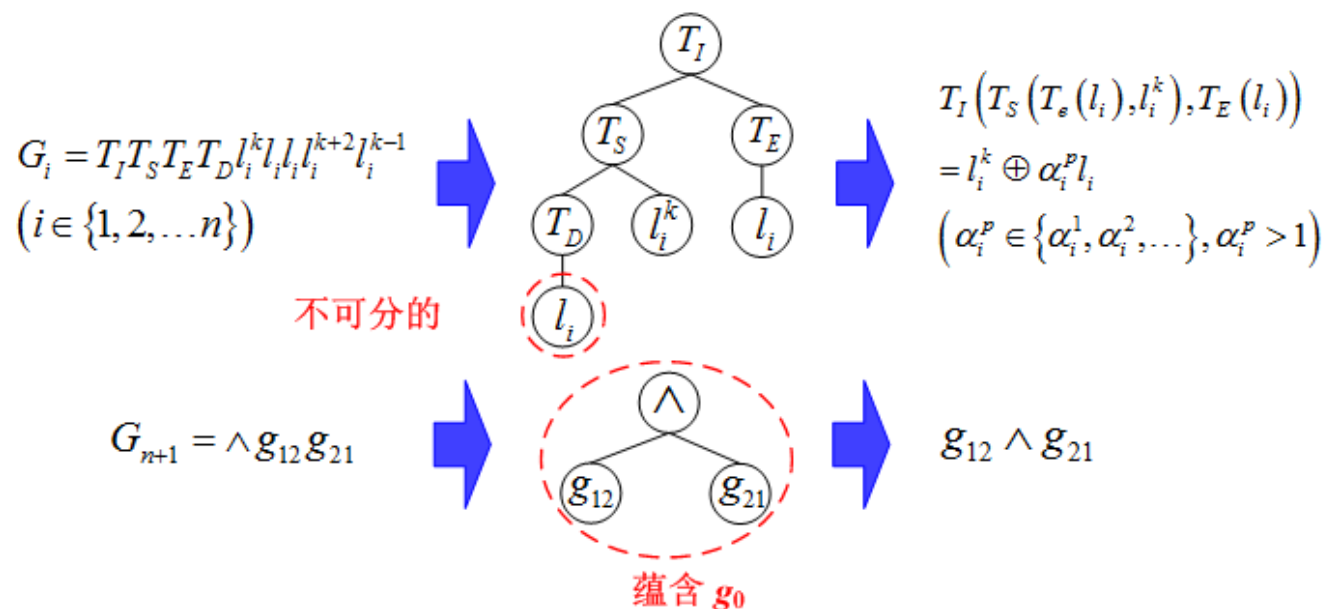
3. 可拓集的应用——可拓策略生成

35

基于GEP的可拓变换运算式自组织构建机制

- 基因序列解码

对可拓变换运算式的解码的独特之处在于，需要考虑每一个可拓变换的合法性。对于叶基元 l_i 基因序列，如果某一个可拓变换与该变换对象 l_i 的拓展推理结果相悖，则该变换被认定为非法，例如，对不可分的基元实施分解变换，对不可加的基元实施增加变换等等。非法的可拓变换将被解码为恒等变换。若基因序列中含有扩缩变换，乘子应根据可扩分析的结果随机进行选取。对于目标 g_0 基因序列，若解码后的目标不蕴含原目标 g_0 ，则应解码为原目标 g_0 。



3. 可拓集的应用——可拓策略生成

36

基于GEP的可拓变换运算式自组织构建机制

进行变异操作生成新的个体，实现在变换空间中的全局搜索

编解码规则+评价准则

实施个体选择机制，对优良个体进行遗传保留，引导可拓变换式的构建过程向好的方向发展

3. 可拓集的应用——可拓策略生成

37

基于GEP的可拓变换运算式自组织构建机制

例：广州自助游策略生成问题

原问题可拓模型

$$G = \begin{bmatrix} \text{游览, 支配对象, 广州} \\ \text{施动对象, 人}D \\ \text{方式, 自助游} \\ \text{时间, } [0,3] \text{ 天} \\ \text{花费, } [0,800] \text{ 元} \end{bmatrix} \cdots$$
$$L = \begin{bmatrix} \text{游览, 支配对象, 经典线路} \\ \text{施动对象, 人}D \\ \text{时间, 4天} \\ \text{花费, 1225.4元} \end{bmatrix}$$

核问题可拓模型

$$g_0 = \begin{bmatrix} \text{路线}Z_{g0}, \text{ 游览所需时间}c_{0r1}, [0,3] \text{ 天} \\ \text{游览所需费用}c_{0r2}, [0,800] \text{ RMB} \end{bmatrix} \cdots$$
$$l_0 = \begin{bmatrix} \text{经典路线}Z_{l0}, \text{ 游览所需时间}c_{0r1}, 4 \text{ days} \\ \text{游览所需费用}c_{0r2}, 1225.4 \text{ RMB} \end{bmatrix}$$

第一天: Chen Clan Academy → (地铁) Chimelong Paradise

酒店: White House Hotel

第二天: Ocean World → (出租车) Olympic Center

酒店: Jingyue Hotel

第三天: Xiangjiang Wild Life Park → (地铁) Citic Plaza

酒店: Xinhua Hotel

第四天: Lianhua Mountain → (出租车) Luhua Park

经典线路

基于GEP的可拓变换运算式自组织构建机制

38

		游览所需时间(h) [↕]	游览所需费用(RMB) [↕]
1 [↕]	Chen Clan Academy [↕]	2 [↕]	10 [↕]
2 [↕]	Chimelong Paradise [↕]	7 [↕]	160 [↕]
3 [↕]	Ocean World [↕]	6 [↕]	100 [↕]
4 [↕]	Olympic Center [↕]	3 [↕]	10 [↕]
5 [↕]	Xiangjiang Wild Life Park [↕]	7 [↕]	120 [↕]
6 [↕]	Citic Plaza [↕]	1 [↕]	0 [↕]
7 [↕]	Lianhua Mountain [↕]	5 [↕]	30 [↕]
8 [↕]	Luhu Park [↕]	3 [↕]	5 [↕]
9 [↕]	Baiyun Mountain [↕]	5 [↕]	30 [↕]
10 [↕]	Yuexiu Park [↕]	3 [↕]	5 [↕]
11 [↕]	Tian He Piao Juan [↕]	2 [↕]	0 [↕]
12 [↕]	Huanghuagang Park [↕]	3.5 [↕]	8 [↕]
13 [↕]	Sun Yat-sen Memorial Hall [↕]	1.5 [↕]	10 [↕]
14 [↕]	Academy of Fine Arts [↕]	3 [↕]	20 [↕]
15 [↕]	Yuntai Park [↕]	3 [↕]	5 [↕]
16 [↕]	Insurrectional Martyr Cemetery Park [↕]	2.5 [↕]	5 [↕]
17 [↕]	The Nineteen Route Army Monument [↕]	2.5 [↕]	5 [↕]
18 [↕]	World Grand Canyon [↕]	5 [↕]	80 [↕]
19 [↕]	Spaceflight Wonder [↕]	4 [↕]	80 [↕]

发散得到的景点

		每晚所需花费 (RMB) [↕]
1 [↕]	White Horse Hotel [↕]	210 [↕]
2 [↕]	Jingyue Hotel [↕]	168 [↕]
3 [↕]	Xinhua Hotel [↕]	200 [↕]
4 [↕]	Chongqing Building Hotel [↕]	150 [↕]
5 [↕]	Haizhu Hotel [↕]	200 [↕]
6 [↕]	Yuxin Hotel [↕]	200 [↕]
7 [↕]	Wenxing Business Hotel [↕]	160 [↕]
8 [↕]	Xuanyi Hotel [↕]	170 [↕]
9 [↕]	Motai Chain Hotel [↕]	208 [↕]
10 [↕]	Home Inns Hotel [↕]	208 [↕]
11 [↕]	Lequn Hotel [↕]	230 [↕]
12 [↕]	Golden Goose Hotel [↕]	
13 [↕]	Dongcheng Hotel [↕]	
14 [↕]	China Grand Hotel [↕]	
15 [↕]	Yuexiu Tianan Hotel [↕]	
16 [↕]	Panyu Xiangjiang Hotel [↕]	

发散得到的酒店

表.3- 景点之间的距离信息以及每种交通工具的可用性信息[↕]

景点对 [↕]	相隔距离(km) [↕]	可用性 [↕]		
		出租车 [↕]	地铁 [↕]	公交 [↕]
(1,2) [↕]	15 [↕]	是 [↕]	是 [↕]	否 [↕]
(1,3) [↕]	6.1 [↕]	是 [↕]	否 [↕]	是 [↕]
(1,4) [↕]	16.5 [↕]	是 [↕]	否 [↕]	是 [↕]

景点之间的距离以及每种交通工具的可用性信息

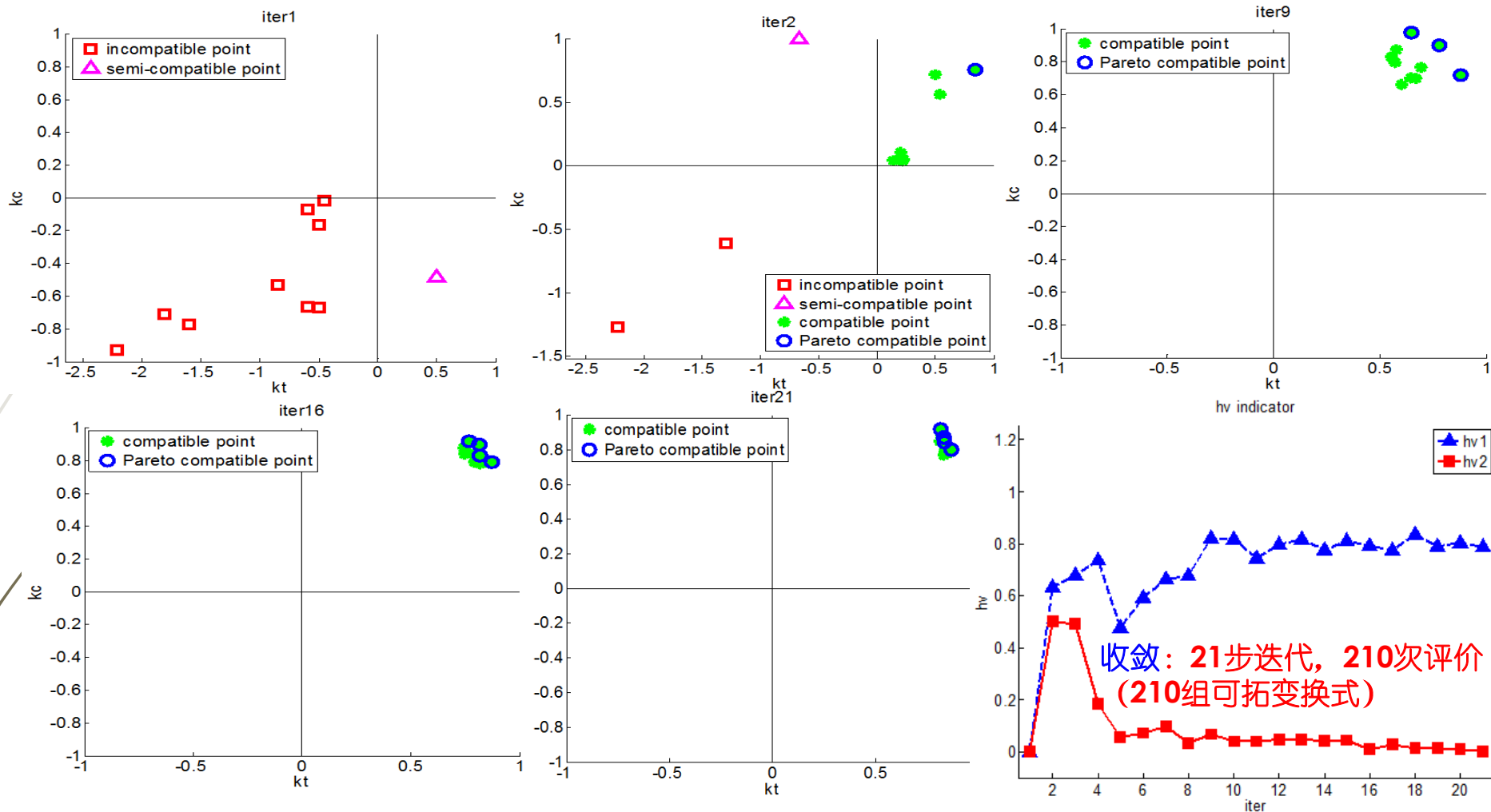
		速度(km/h) [↕]	费用(RMB) [↕]
1 [↕]	地铁 [↕]	110 [↕]	2 (距离不超过 4km) [↕] 2+0.5*(d-4) (距离超过 4km) [↕]
2 [↕]	出租车 [↕]	60 [↕]	7 (距离不超过 2.3km) [↕] 7+2.6*(d-2.3) (距离超过 2.3km) [↕]
3 [↕]	公交 [↕]	20 [↕]	3 [↕]

发散得到的交通工具

$$k_2 = \begin{cases} \frac{x_c}{800}, & x_c \leq 800 \\ \frac{x_c - 800}{800}, & x_c > 800 \end{cases}$$

$$k_1 = \begin{cases} \frac{1}{3} \sum_{k=1}^{n_d} \frac{x_k}{10}, & n_d \leq 3 \\ -\sum_{k=4}^{n_d} \frac{x_k}{10}, & n_d > 3 \end{cases}$$

自助游可拓策略生成



国家自然科学基金项目“基于GEP的可拓策略自组织生成理论与方法研究” (61503085)

Tang Long*, Chunyan Yang, Weihua Li, Adopting GEP to generate extension strategies for incompatible problem. *Neural Computing & Applications*, 2017, 28(9): 2649-2664.

广东工业大学, 签字笔的缺点改进型GEP可拓创新设计软件V1.0, 2019SR0165499, 原始取得, 全部权利, 2018.10.30.

广东工业大学, 关联函数值计算程序软件[简称: 关联函数]V1.0, 2017SR546585, 原始取得, 全部权利, 2017.07.22.

4. 可拓集的应用——可拓数据挖掘

40

基于可拓学的数据挖掘=可拓数据挖掘 ??

数据挖掘：任务（目的）+算法

任务：分类、聚类、关联规则发现或序列模式发现

（分类）算法：决策树、逻辑回归、支持向量机、
人工神经网络、.....

4. 可拓集的应用——可拓数据挖掘

41

- 解决矛盾问题的实质在于通过实施恰当的“变换”来实现问题所涉及对象的改变，从而化解问题；
- 当实施了某个“变换”之后，所涉及的对象会发生怎样的变化？量变还是质变？效果如何？问题是否得到解决？
- **可拓知识：**旨在获取关于对象（特征值、关联度）在变换的作用下所发生的变化（量变、质变）的知识。

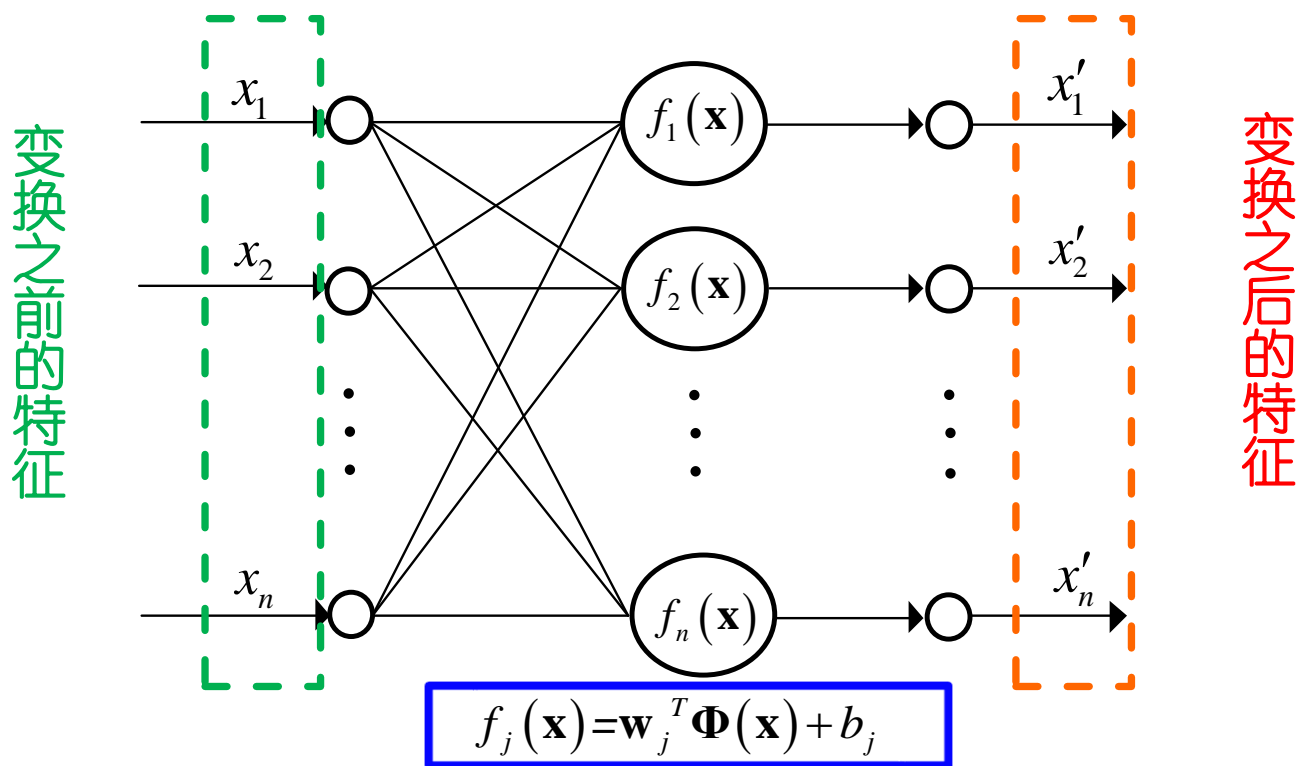
可拓数据挖掘是基于可拓学的数据挖掘的一个分支。

4. 可拓集的应用——可拓数据挖掘

42

可拓支持向量机

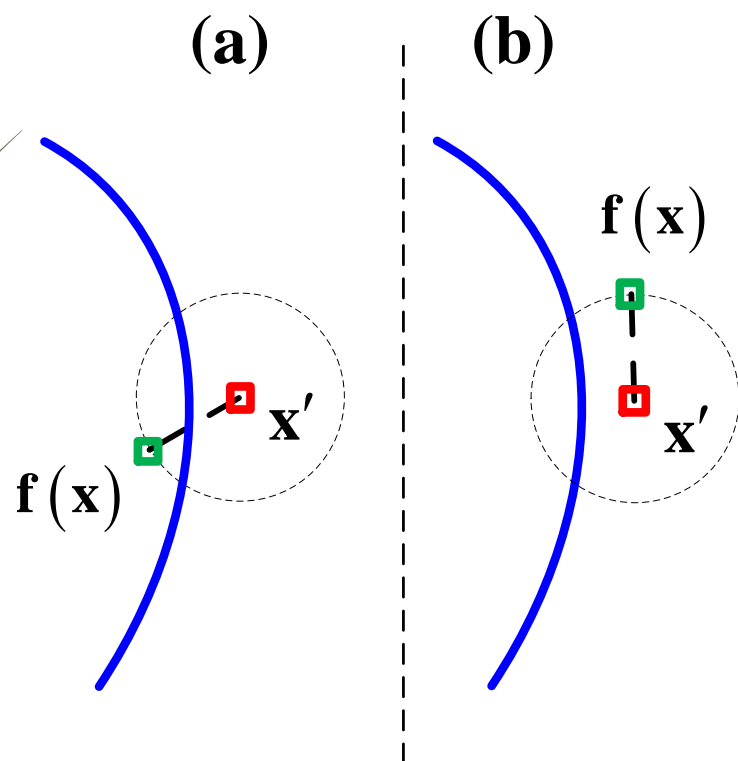
可拓分类：旨在通过挖掘对象在“变换”前后的特征向量之间的数学映射关系，来获取“变换”的作用和效果。



4. 可拓集的应用——可拓数据挖掘

43

可拓支持向量机



$$\min \|f(\mathbf{x}) - \mathbf{x}'\|$$

$$s.t. D(f(\mathbf{x})) D(\mathbf{x}') > 0$$

$$D(\mathbf{x}) = \tilde{\mathbf{w}}^T \tilde{\Phi}(\mathbf{x}) + \tilde{b}$$

4. 可拓集的应用——可拓数据挖掘

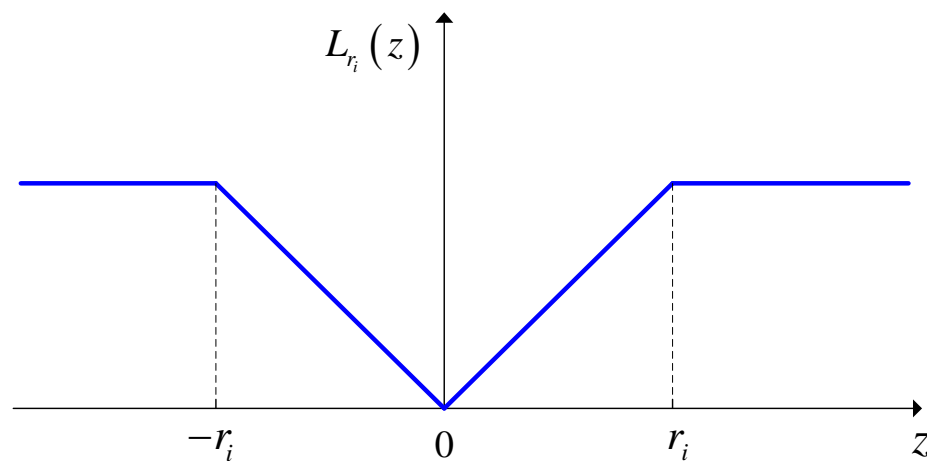
44

可拓支持向量机

$$D(\mathbf{f}(\mathbf{x}))D(\mathbf{x}') > 0$$

的一个充分条件是：

$$\begin{aligned} &|f_j(\mathbf{x}) - x'_j| \leq r \\ &r = \sqrt{-\frac{2}{n} \tilde{\sigma}^2 \ln \left(1 - \frac{D(\mathbf{x}')^2}{2\tilde{\mathbf{W}}^T \tilde{\mathbf{W}}} \right)} - \varepsilon_0 \end{aligned}$$



$$\min_{\mathbf{w}_j, b_j} \frac{1}{2} \mathbf{w}_j^T \mathbf{w}_j + C \sum_{i=1}^N L_{r_i}(f_j(\mathbf{x}^i) - \mathbf{x}''^i)$$

$$L_{r_i}(z) = \begin{cases} |z|, & |z| \leq r_i \\ r_i, & |z| > r_i \end{cases}$$

4. 可拓集的应用——可拓数据挖掘

45

可拓支持向量机

算法 (ESVM) ◀

- 1) 输入训练集 (2) 和关联规则 $\tilde{\mathbf{w}}\tilde{\Phi}(\mathbf{x})+\tilde{b}=\sum_{i=1}^{N_0}\alpha_i^*\tilde{y}^iK(\mathbf{x}^i,\mathbf{x})+\tilde{b}$; ◀
- 2) 利用式 (32) 和 (34), 计算每个训练示例的从属半径; ◀
- 3) 对每个 $j\in\{1,2,\dots,n\}$: ◀

3.1) 初始化迭代 $t=1$, $\mathbf{B}_j^0=(B_{j1}^0,B_{j2}^0,\dots,B_{jn}^0)^T=(0,0,\dots,0)^T$; ◀

3.2) 第 t 次迭代建立并求解 QPP 模型 (52): ◀

$$\begin{aligned} \min_{\alpha_j^t, \alpha_j^{*t}} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^n (\alpha_{ji}^t - \alpha_{ji}^{*t}) (\alpha_{ji}^t - \alpha_{ji}^{*t}) K(\mathbf{x}^i, \mathbf{x}^i) + \sum_{i=1}^N (\alpha_{ji}^t - \alpha_{ji}^{*t}) \mathbf{x}^i \\ \text{s.t.} \quad & \sum_{i=1}^N (\alpha_{ji}^t - \alpha_{ji}^{*t}) = 0 \\ & -B_{ji}^{t-1} \leq \alpha_{ji}^t \leq C - B_{ji}^{t-1}, i=1, \dots, N \\ & 0 \leq \alpha_{ji}^{*t} \leq C, i=1, \dots, N \end{aligned}$$

3.3) 获得: ◀

$$\mathbf{w}_j^t \tilde{\Phi}(\mathbf{x}) = \sum_{i=1}^N (\alpha_{ji}^{*t} - \alpha_{ji}^t) K(\mathbf{x}^i, \mathbf{x}) \dots\dots\dots$$

根据 KKT 条件计算 b_j^t ; ◀

3.4) 对于每个 $i\in\{1,2,\dots,N\}$ 计算: ◀

$$f_j^t(\mathbf{x}^{(i)}) = \mathbf{w}_j^t \tilde{\Phi}(\mathbf{x}) + b_j^t = \sum_{i=1}^N (\alpha_{ji}^{*t} - \alpha_{ji}^t) K(\mathbf{x}^i, \mathbf{x}^i) + b_j^t \dots\dots\dots$$

根据式 (43), 计算: ◀

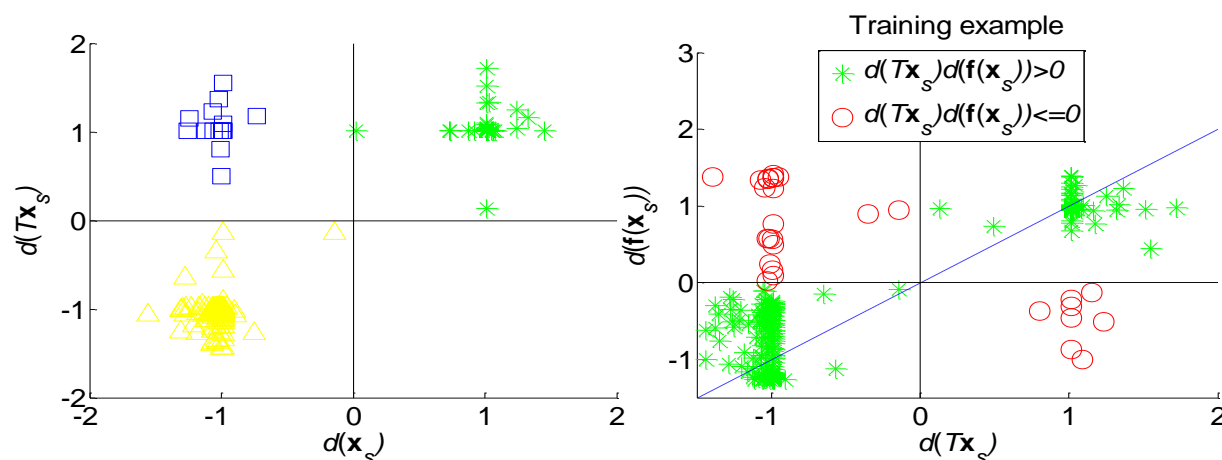
$$V_{ji}^t = \begin{cases} -C, & f_j^t(\mathbf{x}^i) - \mathbf{x}^i < -r_i \\ 0, & -r_i \leq f_j^t(\mathbf{x}^i) - \mathbf{x}^i \leq r_i \\ C, & f_j^t(\mathbf{x}^i) - \mathbf{x}^i > r_i \end{cases} \dots\dots\dots$$

获取 $\mathbf{V}_j^t = (V_{j1}^t, V_{j2}^t, \dots, V_{jn}^t)^T$; ◀

3.5) 如果 $\mathbf{V}_j^t = \mathbf{V}_j^{t-1}$, 收敛; 否则 $t=t+1$, 设置并返回步骤 3.2); ◀

3.6) 获得最终结果 $f_j(\mathbf{x})$ ◀

某通信公司数据集, 涉及300组营销模式转型前后的客户行为数据。考虑了以下三个特性: “呼叫持续时间”、“3G流”和“WLAN流”; 用关联规则 $D(\cdot)$ 描述客户的活跃度。



广东省自然科学基金面上项目“基于GEP高效自组织建模的可拓分类知识获取研究” (2017A030313348)

Long Tang, Chunyan Yang, Xingsen Li, Extension Support Vector Machine for Extension Classification. *Procedia Computer Science* 162 (2019) 174–182 (Best paper award of ITQM2019)

Long Tang*, Xuanbin Lu, Chunyan Yang*, Xingsen Li*, Classification conducting knowledge acquisition by an evolutionary robust GRBF-NN model, *Procedia Computer Science*, 2019, 162: 183–190. (Best paper award of ITQM2019)

4. 可拓集的应用——可拓数据挖掘

46

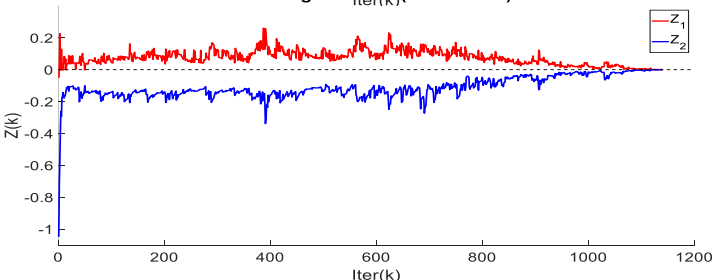
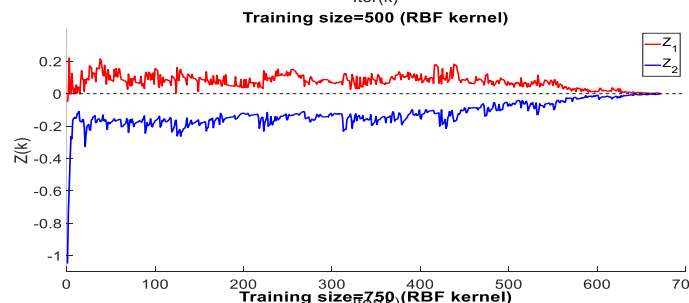
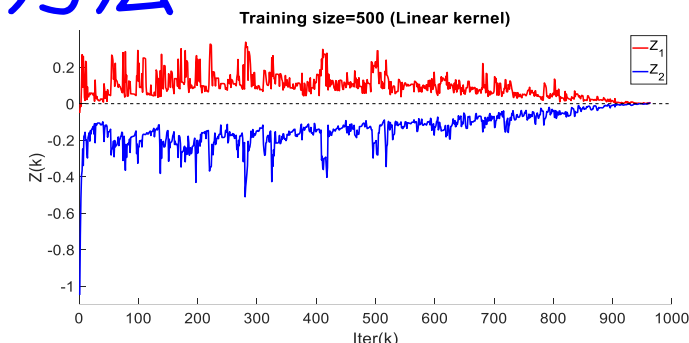
基于经典集和最优化的机器学习方法

Data set.	RSSVM.	IRSSVM.	KSVCR.	SIRSSVM.	1-v-r SNPSVM.	1-v-r SRSVM.
Digits-mfeat-fou_1-5.	90.00±3.16.	90.40±3.58.	90.40±5.18.	90.40±3.58.	90.40±3.58.	84.80±4.60.
Digits-mfeat-fou_6-10.	80.80±5.22.	82.00±2.83.	82.80±3.03.	82.80±4.60.	80.40±2.61.	77.20±3.35.
Digits-mfeat-fac_1-5.	99.60±0.89.	99.60±0.89.	98.80±1.10.	99.60±0.89.	100.00±0.00.	99.20±1.10.
Digits-mfeat-fac_6-10.	98.00±2.00.	98.00±2.00.	97.60±2.61.	98.40±2.61.	98.40±2.19.	98.00±2.00.
Digits-mfeat-kar_1-5.	98.00±1.41.	98.00±1.41.	98.00±1.41.	98.40±0.89.	98.40±0.89.	95.20±2.28.
Digits-mfeat-kar_6-10.	96.40±2.61.	96.40±3.29.	97.20±3.03.	96.80±3.90.	96.80±4.15.	96.00±2.83.
Digits-mfeat-pix_1-5.	98.00±2.45.	98.00±2.45.	97.20±3.03.	98.40±1.67.	98.80±1.79.	95.60±1.67.
Digits-mfeat-pix_6-10.	96.40±1.67.	96.80±2.68.	96.80±2.68.	97.20±1.79.	97.60±1.67.	96.40±2.61.
Digits-mfeat-zer_1-5.	96.80±2.28.	96.40±1.67.	96.80±2.28.	97.20±2.28.	96.80±2.68.	95.20±1.79.
Digits-mfeat-zer_6-10.	77.60±6.07.	76.40±5.73.	76.40±3.85.	77.60±6.54.	76.80±4.60.	74.40±5.18.
Digits-mfeat-mor_1-5.	86.80±3.63.	86.80±4.38.	84.00±4.69.	87.60±3.29.	84.80±5.02.	80.80±2.28.
Digits-mfeat-mor_6-10.	78.80±4.82.	80.00±3.46.	78.00±2.00.	80.80±3.03.	80.80±3.90.	78.40±2.61.
Corel-class_1-3.	77.00±7.85.	78.33±2.04.	77.00±3.21.	79.67±3.42.	78.67±4.62.	74.33±4.94.
Corel-class_4-6.	75.00±5.65.	75.00±5.65.	73.33±5.14.	75.00±5.65.	76.00±4.50.	73.00±6.28.
Corel-class_7-9.	78.33±5.27.	77.33±8.13.	80.33±6.28.	78.67±6.39.	79.33±7.96.	76.67±7.73.
Corel-class_10-12.	72.33±3.65.	72.33±5.48.	73.67±3.21.	73.33±3.12.	73.33±3.12.	74.33±3.25.
Corel-class_13-15.	69.67±7.40.	69.67±3.80.	68.67±7.67.	70.00±5.40.	69.67±4.77.	70.00±4.41.
Corel-class_16-18.	74.67±11.93.	73.67±9.60.	73.67±9.89.	75.33±10.17.	76.00±8.05.	71.67±7.55.
Corel-class_19-21.	88.00±6.39.	88.00±4.62.	88.67±4.62.	88.33±4.71.	88.00±6.60.	86.33±3.61.
Corel-class_22-24.	69.33±7.42.	67.00±8.93.	69.00±6.41.	69.33±7.32.	68.33±8.16.	66.67±3.54.
Corel-class_25-27.	77.00±2.17.	77.00±5.06.	75.67±5.35.	78.00±6.17.	76.67±5.14.	76.00±1.90.
Corel-class_28-30.	94.00±5.35.	94.00±5.35.	94.00±6.73.	94.00±5.35.	95.00±5.53.	94.00±4.50.
Corel-class_31-33.	92.67±1.90.	92.67±1.90.	92.33±3.03.	93.33±1.18.	92.00±2.47.	92.33±2.24.
Corel-class_34-36.	78.00±4.15.	76.67±3.12.	77.67±3.25.	78.67±5.19.	78.67±5.45.	75.33±2.98.
Corel-class_37-39.	86.00±7.32.	86.00±3.46.	85.67±4.94.	86.67±6.35.	86.00±4.50.	85.00±3.54.
Corel-class_40-42.	79.67±3.21.	80.00±3.33.	78.67±9.01.	80.33±2.98.	79.67±4.62.	76.00±6.30.
Corel-class_43-45.	79.67±12.55.	80.67±12.05.	79.67±12.99.	81.00±11.70.	79.00±9.90.	80.00±12.30.
Corel-class_46-48.	91.33±2.74.	91.00±2.53.	90.67±2.53.	91.67±3.54.	92.33±3.25.	90.33±4.47.
Corel-class_49-51.	72.00±3.61.	72.00±3.61.	71.00±2.24.	72.00±3.61.	71.33±1.39.	69.67±3.98.
Corel-class_52-54.	80.33±6.17.	80.33±6.17.	79.33±5.48.	80.33±6.60.	81.33±6.71.	78.67±6.50.
Corel-class_55-57.	87.67±9.17.	88.67±7.01.	87.00±8.28.	89.33±7.69.	89.00±9.90.	87.33±10.04.
Corel-class_58-60.	68.33±9.86.	68.67±10.17.	66.00±10.38.	69.00±10.18.	68.67±12.88.	65.00±8.25.
Corel-class_61-63.	97.00±1.83.	97.00±2.17.	97.33±1.49.	98.33±1.18.	98.33±1.18.	98.00±2.17.
Corel-class_64-66.	86.00±6.52.	87.00±5.58.	84.33±4.50.	89.00±3.46.	88.33±3.54.	85.33±5.45.
Corel-class_67-69.	64.00±6.52.	65.00±5.89.	63.33±3.54.	65.67±8.55.	65.00±7.73.	63.33±5.89.
Corel-class_70-72.	89.33±3.03.	90.00±3.91.	89.33±1.90.	90.00±3.91.	89.67±2.98.	88.67±3.42.
Corel-class_73-75.	75.00±7.73.	76.33±9.75.	73.33±8.98.	76.33±9.82.	76.67±8.74.	71.67±11.79.
Corel-class_76-78.	93.67±3.21.	93.67±3.21.	94.00±3.84.	93.67±2.98.	94.00±3.03.	93.67±3.61.

$$\min_{\mathbf{w}, \mathbf{b}} \sum_{j=1}^{K-1} \frac{1}{2} (\mathbf{w}_j^T \mathbf{w}_j + b_j^2) + C \sum_{i=1}^N \sum_{k \neq c_i} \xi_{i,k}$$

$$\text{s.t. } \sum_{j=1}^{K-1} (2(V_{c_i,j} - V_{k,j})(\mathbf{w}_j^T \mathbf{x}_i + b_j) + V_{k,j}^2 - V_{c_i,j}^2) \geq \varepsilon - \xi_{i,k}, (i = 1, 2, \dots, N)$$

$$\xi_{i,k} \geq 0, (i = 1, 2, \dots, N)$$



4. 可拓集的应用——可拓数据挖掘

47

基于经典集和最优化的机器学习方法

1. Long Tang*, Yingjie Tian, Wenjun Li, Panos M. Pardalos*. Valley-loss regular simplex support vector machine for robust multiclass classification, *Knowledge-based systems*, 2021, 216, 106801. (SCI, IF=5.921)
2. Long Tang, Yingjie Tian, Wenjun Li, Panos M. Pardalos*. Structural improved regular simplex support vector machine for multiclass classification, *Applied soft computing*, 2020, 91, 106235. (SCI, IF=5.472)
3. Long Tang, Yingjie Tian, Panos M. Pardalos*. A novel perspective on multiclass classification: regular simplex support vector machine. *Information sciences*, 2019, 480: 324-338. (SCI, IF=5.910)
4. Long Tang, Yingjie Tian*, Chunyan Yang. Nonparallel support vector regression model and its SMO-type solver. *Neural networks*, 2018, 105: 431-446. (SCI, IF=5.535)
5. Long Tang, Yingjie Tian, Chunyan Yang, Pan. M Pardalos*. Ramp-loss nonparallel support vector regression: robust, sparse and scalable approximation. *Knowledge-based systems*, 2018, 147: 55-67. (SCI, IF=5.921)

探索：基于可拓集和最优化的机器学习模型和方法。

感谢各位聆听！