因素空间理论下基点分类算法研究

蒲凌杰， 曾繁慧， 汪培庄

引用本文:
蒲凌杰, 曾繁慧, 汪培庄. 因素空间理论下基点分类算法研究[J]. 智能系统学报, 2020, 15(3): 528–536.
PU Lingjie, ZENG Fanhui, WANG Peizhuang. Base point classification algorithm based on factor space theory[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(3): 528–536.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201903031

DOI: 10.11992/tis.201903031

因素空间理论下基点分类算法研究

蒲凌杰[1,2], 曾繁慧[1,2], 汪培庄[1,2]

(1. 辽宁工程技术大学 理学院, 辽宁 阜新 123000; 2. 辽宁工程技术大学 智能工程与数学研究院, 辽宁 阜新 123000)

摘 要: 目前, 基于因素空间理论的背景基提取算法计算过程复杂, 初始化必须依赖各因素极值, 基点数量提取冗余等原因, 未能在应用中取得很好效果。为此, 结合内点判别法和知识可继承、可扩展的思想, 提出一种计算简单、初始化独立、基点数量小的改进的背景基提取算法。然后, 利用改进的背景基提取算法构造出一种全新的数据分类算法——基点分类算法, 基点分类算法以提取每一类样本的背景基为预测模型, 再通过新定义的 λ-背景基, 优化预测模型。数值实验表明: 基点分类算法原理简单、构造难度小、分类模型泛化能力强, 预测能力准确率高, 同时严格的模型限定区域又能为识别新类别提供新方法。

关键词: 因素空间; 背景基; 背景基提取; λ-背景基; 基点分类算法; 识别新类别; 数据分类; 背景分布; 背景关系
中图分类号: TP18 **文献标志码:** A **文章编号:** 1673−4785(2020)03−0528−09

中文引用格式: 蒲凌杰, 曾繁慧, 汪培庄. 因素空间理论下基点分类算法研究 [J]. 智能系统学报, 2020, 15(3): 528–536.
英文引用格式: PU Lingjie, ZENG Fanhui, WANG Peizhuang. Base point classification algorithm based on factor space theory[J]. CAAI transactions on intelligent systems, 2020, 15(3): 528–536.

Base point classification algorithm based on factor space theory

PU Lingjie[1,2], ZENG Fanhui[1,2], WANG Peizhuang[1,2]

(1. College of Science, Liaoning Technical University, Fuxin 123000, China; 2. College of Intelligent Engineering and Mathematics, Liaoning Technical University, Fuxin 123000, China)

Abstract: At present, the background-based extraction algorithm based on factor space theory has not achieved good results when used in applications. Reasons for its inefficiency are the calculation process is complicated, initialization depends on the extreme values of each factor, and redundancy of the number of base points extracted. Therefore, combining the inner point judgment method and a novel idea, an improved background-based extraction algorithm with simple calculation, independent initialization, and a few number of base points is proposed. Using the improved background-based extraction algorithm, a new data classification algorithm, i.e., base point classification algorithm, is constructed. The algorithm extracts the background base of each type of sample as the prediction model and optimizes the prediction model through the newly defined λ-background base. Finally, numerical experiments show that the base point classification algorithm is simple in principle, easy in construction, strong in generalizing the ability of classification model, and high in accuracy of prediction ability. Moreover, strict-utility model can provide new methods for identifying new classes.

Keywords: factor space; background base; background base extraction; λ-background base; base point classification algorithm; identify new classes; data classification; background distribution; background relationship

机制主义人工智能理论框架包括: 机制主义人工智能理论[1]、泛逻辑学理论[2]和因素空间理论[3]。因素空间理论是机制主义人工智能的数学基础。背景基理论是因素空间理论框架下的一个重要分支。因素空间理论是现有人工智能数学理论的进一步提升, 该理论从形成事物的本质出发, 以构成事物的“基因”为维度, 形成一个描述

收稿日期: 2019−03−23.
基金项目: 国家自然科学基金委主任基金项目 (61350003); 辽宁省教育厅科学技术研究经费项目 (LJ2019JL019).
通信作者: 蒲凌杰. E-mail: 1901676469@qq.com.

事物的“基因空间”。基因最早的英文名称是 Factor, 因素就是广义的基因, 所以用“因素空间”描述事物更具有普适价值。背景分布是因素空间理论下的一个重要概念。2013 年, 汪培庄[4] 在讨论因素空间与因素库的关系时提出背景分布概念。2014—2015 年, 汪培庄[5-6] 在因素空间与数据科学一文中讨论了数据科学与因素背景关系的深刻联系。“背景分布–背景关系–背景基”模型是人对客观事物从观察到认识事物, 再到知识形成的过程[7-9]。背景基的提出就是为了描述因素的背景关系。曲国华等[10-11] 研究了背景分布与模糊背景之间的关系。与此同时, 汪培庄提出了一种构造背景基的判别方法: 内点判别法。该算法简单、高效, 可以近似描述背景关系。吕金辉等[12] 为了弥补该算法的近似性, 给出了完整的夹角判别定理, 使得内点判别法成为一种精确的算法, 但该方法计算较复杂, 且在高维空间中提取的基点较冗余。此外, 目前大数据计算复杂性理论仍然需要时间突破[13]。跳出时间复杂度成指数增长的约束仍是专家学者们继续研究的热点。针对以上问题, 提出了一种基于背景基的数据分类算法: 基点分类算法。该算法可以模仿人类认识事物的过程, 把人类学习到的知识用背景基刻画, 不同的背景基容纳不同知识体, 最终实现知识的分类与预测。从机制主义人工智能理论的角度来看, 该算法不但有监督学习和预测的能力, 而且在遇到全新知识时, 有自我判断和自我分类的能力, 所以说, 该算法从原理上可以认为是具有一定的智能生长能力[14]。本文提出可继承可扩展的改进的背景基提取算法 (improved background-base extraction, IBBE), 设计基点分类算法, 通过定义 λ-背景基优化预测模型, 数值实验验证算法有效性。

1　背景基理论基本概念

因素空间理论中, 认为因素是事物的质根, 强调事物构成的根本。背景基则强调用集合理论表达知识

定义 1　一类事物的集合称为论域, 用 U 表示。

定义 2　映射。$f: U \to X(f)$称为一个因素, 其中 $X(f)$ 是 f 从事物所映射出来的性状的集合, 称为 f 的性状空间。

对于给定论域 U 上的一组因素 $f_j: U \to X(f_j)$ $(j=1,2,\cdots,n)$, 称集合 $F^*=\{f_1,f_2,\cdots,f_n\}$ 为当前论域 U 的一个因素集。$P(F^*)$ 为 F^*的幂集, 元素个数为 $2^{|F^*|}$, 此处 $|F^*|=n$, $P(F^*)$ 的任意元素 $F_i=\{f_{(j)},\cdots,f_{(k)}\}$, 其中 $i=1,2,\cdots,2^{|F^*|}$, $\{f_{(j)},\cdots,f_{(k)}\}$ 是 $P(F^*)$ 的一个子集合, 定义一个 U 上的合成因素 F_i: $U \to X(F_i)$, 其性状空间是

$$X(F_i)=X(f_{(j)})\times\cdots\times X(f_{(k)}) \tag{1}$$

记为 $F_i=f_{(j)}\cup\cdots\cup f_{(k)}$。式 (1) 的含义是: 合成因素 F_i 的性状空间 $X(F_i)$ 被定义成一个笛卡尔乘积 $X(f_{(j)})\times\cdots\times X(f_{(k)})$。

定义 3　因素空间。记 $X_{F^*}=\{X(F)\}_{(F\in P(F^*))}$, 称 $\Phi=(U,X_{F^*})$ 为 U 上的一个因素空间。

记 $X_{\max}=X(f_1)\times X(f_2)\times\cdots\times X(f_n)$称为最大性状空间。对于离散型性状空间而言, 任意 $\boldsymbol{a}=(a_1,a_2,\cdots,a_n)\in X$ 称为一个性状颗粒。

因素空间定义的更多解释, 可参考文献 [3] 中第 1 章的相关内容。

定义 4　背景关系。给定 U 上的定性因素空间 $\Phi=(U,X_{F^*})$, 对任意相 $\boldsymbol{a}=(a_1,a_2,\cdots,a_n)\in X$, 记其在 U 上的原相为

$$[\boldsymbol{a}]=\boldsymbol{F}^{-1}(\boldsymbol{a})=\{u\in U|\boldsymbol{F}(u)=\boldsymbol{a}\} \tag{2}$$

$[\boldsymbol{a}]$ 可能是空集, 若 $[\boldsymbol{a}]\neq\varnothing$, 则称 $\boldsymbol{a}$ 是一个实性状颗粒, 否则称 $\boldsymbol{a}$ 是一个虚组态。全体实性状集合记为

$$\begin{aligned}R=\boldsymbol{F}^*(U)=\{&\boldsymbol{a}=(a_1,a_2,\cdots,a_n)\in X|\exists u\in U\\ &f_1(u)=a_1,f_2(u)=a_2,\cdots,f_n(u)=a_n\},\end{aligned} \tag{3}$$

R 称为因素 $f_1,f_2,\cdots,f_n$ 之间的背景关系, 也称为因素 F^*的背景集。

机器学习中 $\boldsymbol{a}$ 可以视为一条样本, $[\boldsymbol{a}]$ 表示 $\boldsymbol{a}$ 的对应标签。如果一个 $\boldsymbol{a}$ 有对应标签, 则 $\boldsymbol{a}$ 是有意义的, 即 $\boldsymbol{a}$ 是一个实性状颗粒, 否则 $\boldsymbol{a}$ 是没有意义的, 称 $\boldsymbol{a}$ 是一个虚组态。背景关系可以理解为: 所有有意义的样本构成的集合 S 与样本的标签之间的因果关系。若要深入讨论集合 S 的分布问题, 则引出定义 5。

定义 5　背景分布。设论域 $U=(U,\boldsymbol{A},p)$ 是一个概率场, $\Phi=(U,X_{F^*})$ 是定义在 U 上的一个因素空间, $X=(X,\boldsymbol{B})$ 是最大性状空间上的一个可测结构。若所有 F^* 中的 f_j 都是可测映射, 记 $p=p_{F^*}$ 为 p 经过 F^* 在 X 上所诱导出来的概率, 亦即对任意 $B\in\boldsymbol{B}$, 都有 $p(B)=p\left((F^*)^{-1}(B)\right)$。$B$ 称为因素 F^* 的背景分布。

定义 6　背景基。若每个性状空间 $X(f_j)$ 都是有序集, 背景关系 R 是 X 中的凸集, 记 R 的所有顶点所构成的集合为 B=$B(R)$=$\{P|P$ 是 R 的顶点$\}$, 称作背景基。S 表示样本集合, 当 S=R 时, 记 B 的所有顶点构成的集合为 $B(S)$=$\{P|P$ 是 S 的顶点$\}$, 称为样本背景基。

背景基可以生成背景关系, 是背景关系的无

信息损失的压缩，对因素库[4]的实际应用具有重要的意义。

为了更好地说明背景关系、背景分布、背景基三者概念之间的联系，以图 1 为例进行图示说明。

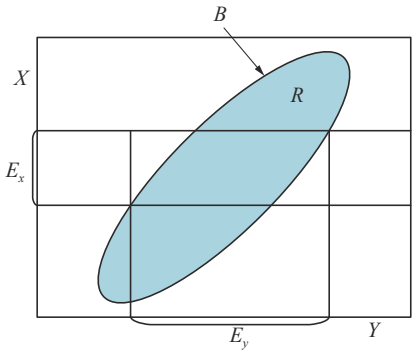

图 1　“背景分布−背景关系−背景基”关系

Fig. 1　“Background distribution-background relationship-background base” relationship

设集合 X、Y、R，其中 $E_x \subseteq X, E_y \subseteq Y$，集合 R 称为背景关系，集合 R 的“边缘”构成背景基 B，p 是背景分布，视为集合 R 的概率密度，可以由样本的相 $\boldsymbol{a}$ 的频率逼近来实现。背景关系 R 表示集合 X 和集合 Y 之间的关联性。这里有

$$E_x \, R \, E_y$$

可得逻辑推理：

$$E_x \rightarrow R_y(R\text{表示} \rightarrow)$$

因此可以认为，背景关系 R 保证了 X 与 Y 的因果关系，背景基则限定了背景关系区域，背景分布表示了因果关系的概率。

2　改进的背景基提取算法

集合论的诞生，使得数学与认知实现了潜在的融合。人类对事物的认知表现可以用数学中的集合刻画，集合可以表现概念，认知的逻辑特性被集合论充分地表现出来。因素空间的背景基理论就是力求用数学刻画人的学习过程。从技术层面来讲，目前由夹角判别定理实现的知识归纳算法有背景基提取算法[12]，但是该算法判别条件复杂、时间复杂度较大，搜寻到的基点数量太多。这些导致表达的知识不利于继承和保存。为此，本章对现有算法改进，使其在减少复杂度同时，更符合人的思维逻辑。

2.1　背景基提取原理

用技术手段实现背景基的提取，是“背景关系−背景分布−背景基”从理论到实践的关键。目前背景基提取主要用到夹角判别定理。

夹角判别定理[12]　设 S 是一个样本集。S^* 是 S 的一个基点集合，其中心为 O。$P \in X \backslash S^*$ 是一个新的样本点。它可被删除当且仅当存在一点 $Q \in S^*$，使射线 PQ 与射线 PO 形成钝角，亦即 $(Q-P, O-P)<0$。而且，对于任意 $Q' \in S^* \backslash \{Q\}$，都有

$$(P-O, O^*-O) \leqslant (O^*-O, O^*-O)$$

这里，O^* 是 O 在直线 QQ' 上的垂足。

内点判别法　设 S 为数据样本集，P 是 S 的一个内点当且仅当在 S 中存在一点 Q，使射线 PQ 与射线 PO 形成钝角，亦即 $(Q-P, O-P)<0$。

夹角判别定理是一种精确求取背景基的方法，但该方法在计算时较复杂，在高维空间提取的基点较冗余。为此，在简化计算的同时，减少背景基点个数，同时保证信息不损失是本章算法改进的目的。

内点判别法计算简单，基点提取率高效，有利于处理大样本高维度数据。为此，本章采用内点判别法作为改进背景基提取算法的理论基础。

2.2　改进的背景基提取算法

给定一组样本集合 $S=\{\boldsymbol{x}_i=(x_{i1}, x_{i2}, \cdots, x_{in})\}$，$n$ 表示样本个数，$|S|$ 表示样本数量，$|S|=m$，$i=1,2,\cdots,m$，且 $m>n$。设 $\boldsymbol{B}$ 为背景基，初始值 $\boldsymbol{B}=\varnothing$，含义表示人对事物的认知初始处于混沌状态，随着对样本的学习，$\boldsymbol{B}\neq\varnothing$，表示人对事物开始有了一定的判断力，随着样本量的增加，$\boldsymbol{B}$ 的泛化能力越来越强，当样本足够充分时，就会形成对特定事物的认知包，一个认知包就是一个知识，或者一个概念。

改进的背景基提取算法就是在模仿人学习过程同时提取知识轮廓，保存学到的知识。下面给出改进的背景基提取算法步骤。

算法　改进的背景基提取算法 (IBBE 算法)。

输入　样本 $S=\{x_i=(x_{i1}, x_{i2}, \cdots, x_{in})\}$。

初始化　背景基 $B=\varnothing$；样本个数 $|S|$；计数器 count=0；θ 为初始基点个数。

1) 取 S 中样本 $\boldsymbol{x}_i$，若 $|B|<\theta$，则 $x_i \in B$，否则，转至 2)。

2) 计算 B 的几何中心 $\boldsymbol{o}$：

$$\boldsymbol{o}=\frac{1}{k}\sum_{j=1}^{k}\boldsymbol{b}_j, \quad \boldsymbol{b}_j \in \boldsymbol{B}$$

计算内积：

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}_j)=\boldsymbol{\alpha}\boldsymbol{\beta}_j^{\mathrm{T}}$$

其中：$\boldsymbol{\alpha}=\boldsymbol{o}-\boldsymbol{x}_i$，$\boldsymbol{\beta}_j=\boldsymbol{b}_j-\boldsymbol{x}_i$，$j=1,2,\cdots,|B|$。

若存在 $\boldsymbol{\beta}_j$ 使得 $(\boldsymbol{\alpha}, \boldsymbol{\beta}_j)<0$，那么 $\boldsymbol{x}_i$ 是背景基 B 的内点，舍去，并转至 1)；否则 $\boldsymbol{x}_i$ 是 B 的一个基点，即 $\boldsymbol{x}_i \in B$，然后转至 3)。

3) 在 2) 中加入 $\boldsymbol{x}_i$ 可能会导致一个或多个基

点再次变成内点, 而内点需要被剔除, 因此 3) 功能是剔除 B 中内点。具体如下:

①取基点 $\boldsymbol{b}_j \in \boldsymbol{B}$ 视为待测点, $\boldsymbol{B}$ 中剩余的基点 $B^* = B\backslash \boldsymbol{b}_j$ 视为临时背景基; 然后计算 B^*的几何中心 $\boldsymbol{o}'$ 以及内积:

$$(\boldsymbol{\alpha}', \boldsymbol{\beta}_i') = \boldsymbol{\alpha}'(\boldsymbol{\beta}_i')^{\mathrm{T}},$$

其中: $\boldsymbol{\alpha}' = \boldsymbol{o}' - \boldsymbol{b}_j$, $\boldsymbol{\beta}_i' = \boldsymbol{b}_i^* - \boldsymbol{b}_j$, $\boldsymbol{b}^* \in B^*$。

②若 $(\boldsymbol{\alpha}', \boldsymbol{\beta}_i') < 0$, 则继续判断 b_j 是否为极点, 即 $\boldsymbol{b}_j = P_j^+$? 或 $\boldsymbol{b}_j = P_j^-$? 其中

$$P_j^+ = \mathrm{argmax}_i\{b_{ij} | b_{ij} \in \boldsymbol{B}\}$$

$$P_j^- = \mathrm{argmin}_i\{b_{ij} | b_{ij} \in B\}(j = 1,2,\cdots,|\boldsymbol{B}|)$$

如果不是各因素的极点, 则标记 $\boldsymbol{b}_i$ 为待删除点 d_i; 依此类推, 检测 $\boldsymbol{B}$ 中所有基点。

4) 若 $\boldsymbol{B}$ 中存在内点, 则删除标记点 d_i 标记的基点, 得到 $\boldsymbol{B}'$, 更新 $\boldsymbol{B}$: $\boldsymbol{B} \leftarrow \boldsymbol{B}'$; 再返回 1) 读取全部样本后转至 5)。

5) 输出背景基 $\boldsymbol{B}$。

通常背景基 $\boldsymbol{B}$ 是由矩阵存储, 若样本 S 有 n 个因素, $\boldsymbol{B}$ 的基点个数$|\boldsymbol{B}|$=m, 则 $\boldsymbol{B}$ 是一个 $m\times n$ 的矩阵:

$$\boldsymbol{B}_{m\times n} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

$\boldsymbol{B}$ 的每一行表示一个基点。

2.3　算例

图 2 中, 样本集合 S 包含点 $\boldsymbol{a}$=(1,3)、$\boldsymbol{b}$=(1,1)、$\boldsymbol{c}$=(2,4)、$\boldsymbol{d}$=(2,3)、$\boldsymbol{e}$=(4,2) 和 $\boldsymbol{f}$=(3,1), 利用上述改进算法生成样本集合 S 的背景基 $\boldsymbol{B}$。

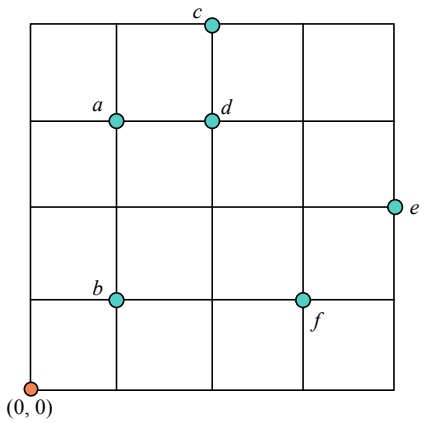

图 2　背景基内点判别法

Fig. 2　Angle criterion for the inner points of the background base

计算步骤如下:

输入　样本集合 S={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{e}$, $\boldsymbol{f}$}。

初始化　背景基 $\boldsymbol{B}$ = Ø; 样本个数$|S|$=6; 计数器 count=0。

选代 t=1:

1) 从 S 中读取一个样本 $\boldsymbol{a}$, count=1, 此时$|\boldsymbol{B}|$=1, $|\boldsymbol{B}|$<3, 不做任何判断, 此时 $\boldsymbol{B}$={$\boldsymbol{a}$};

从 S 中取样本 $\boldsymbol{b}$, count=2, 此时$|\boldsymbol{B}|$=2, $|\boldsymbol{B}|$<3,不做任何判断, 此时 $\boldsymbol{B}$={$\boldsymbol{a}$, $\boldsymbol{b}$};

从 S 中取样本 $\boldsymbol{c}$, count=3, 此时$|\boldsymbol{B}|$=2, $|\boldsymbol{B}|$<3, 条件不成立, 转至 2)。

2) 先计算 $\boldsymbol{B}$ 的中心:

$$\boldsymbol{o} = \frac{1}{k}\sum_{j=1}^{k} b_j = \frac{1}{2}(a+b) = (1,2)$$

再计算内积:

$(\boldsymbol{o}-\boldsymbol{c}, \boldsymbol{a}-\boldsymbol{c})$=$(\boldsymbol{o}-\boldsymbol{c})(\boldsymbol{a}-\boldsymbol{c})^{\mathrm{T}}$=$(-1,-2)(-1,-1)^{\mathrm{T}}$=3>0

$(\boldsymbol{o}-\boldsymbol{c}, \boldsymbol{b}-\boldsymbol{c})$=$(\boldsymbol{o}-\boldsymbol{c})(\boldsymbol{b}-\boldsymbol{c})^{\mathrm{T}}$=$(-1,-2)(-1,-3)^{\mathrm{T}}$=7>0

因为上述内积结果都为正, 即交成锐角, 所以认为 $\boldsymbol{c}$ 是一个基点, 即 $\boldsymbol{B}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$}。

3) 在 2) 中 $\boldsymbol{B}$ 的更新可能产生新的内点, 需要进一步筛选和剔除:

①将 $\boldsymbol{B}$ 中的 $\boldsymbol{a}$ 视为内点, $\boldsymbol{B}^*$=$\boldsymbol{B}\backslash\boldsymbol{a}$={$\boldsymbol{b}$,$\boldsymbol{c}$}, 按照 2) 方法判断, 得到 $\boldsymbol{a}$ 不是 $\boldsymbol{B}^*$中的内点;

②将 $\boldsymbol{B}$ 中的 $\boldsymbol{b}$ 视为内点, $\boldsymbol{B}^*$=$\boldsymbol{B}\backslash\boldsymbol{b}$={$\boldsymbol{a}$,$\boldsymbol{c}$}, 按照 2) 方法判断, 得到 $\boldsymbol{b}$ 不是 $\boldsymbol{B}^*$中内点; 转至 4)。

4) $\boldsymbol{B}$ 中没有内点, 故 $\boldsymbol{B}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$}; 此时 count=3<$|S|$, 则转至 1)。

选代 t=2:

1) 从 S 中取样本 $\boldsymbol{d}$, count=4, 此时$|\boldsymbol{B}|$=3, $|\boldsymbol{B}|$<2, 条件不成立, 转至 2)。

2)~4) 与上述对应步骤相同, 此处省略, 得到 $\boldsymbol{B}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$}。

选代 t=3:

1) 从 S 中取样本 $\boldsymbol{e}$, count=5, 此时$|\boldsymbol{B}|$=4, $|\boldsymbol{B}|$<2, 条件不成立, 转至 2)。

2) 与上述对应步骤相同, 此处省略, 得到 $\boldsymbol{B}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{e}$}。

3) ①取 $\boldsymbol{B}$ 中 $\boldsymbol{a}$ 为待测点, $\boldsymbol{B}^*$=$\boldsymbol{B}\backslash\boldsymbol{a}$={$\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{e}$}, 判断得到 $\boldsymbol{a}$ 不是 $\boldsymbol{B}^*$中内点;

②取 $\boldsymbol{B}$ 中 $\boldsymbol{b}$ 为待测点, $\boldsymbol{B}^*$=$\boldsymbol{B}\backslash\boldsymbol{b}$={$\boldsymbol{a}$, $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{e}$}, 判断得到 $\boldsymbol{b}$ 不是 $\boldsymbol{B}^*$中内点;

③取 $\boldsymbol{B}$ 中 $\boldsymbol{c}$ 为待测点, $\boldsymbol{B}^*$=$\boldsymbol{B}\backslash\boldsymbol{b}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{d}$, $\boldsymbol{e}$}, 判断得到 $\boldsymbol{c}$ 不是 $\boldsymbol{B}^*$中内点;

④取 $\boldsymbol{B}$ 中 $\boldsymbol{d}$ 为待测点, $\boldsymbol{B}^*$=$\boldsymbol{B}\backslash\boldsymbol{b}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{e}$}, 判断得到 $\boldsymbol{d}$ 是 $\boldsymbol{B}^*$中的内点, 再根据极点公式 P_j^+、P_j^- 判断得到 $\boldsymbol{d}$ 不是极点, 因此标记 $\boldsymbol{d}$ 点为待删除点 d_1。

4) $\boldsymbol{B}$ 中有内点存在, 删除标号 d_1 的点, 所以 $\boldsymbol{B}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{e}$}; 此时 count=5<$|S|$, 则转至 1)。

选代 t=4:

1) 从 S 中取样本 $\boldsymbol{f}$, count=6, 此时$|\boldsymbol{B}|$=4, $|\boldsymbol{B}|$<2, 条件不成立, 转至 2)。

2)、3) 与上述对应步骤相同, 此处省略, 得到 B={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{e}$, $\boldsymbol{f}$}。

4) $\boldsymbol{B}$ 中没有内点, 故 $\boldsymbol{B}$={$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{e}$, $\boldsymbol{f}$}; 此时 count=6=$|S|$, 循环结束。

输出　背景基 $\boldsymbol{B}$。

输出的背景基 $\boldsymbol{B}$ 是 5×2 矩阵, 为表达上的美观, 采用 $\boldsymbol{B}$ 的转置表达, 即

$$\boldsymbol{B}=\begin{bmatrix}1&1&2&4&3\\3&1&4&2&1\end{bmatrix}^{\mathrm{T}}$$

$\boldsymbol{B}$ 的每列代表一个因素, 如图 3 所示, 虚线包含的区域可以近似视为由样本数据生成的背景分布, 此时背景基为背景分布轮廓的特征点集合。由上述算例可得背景基在信息压缩、知识表达与存储等方面有着独具创新的优势。

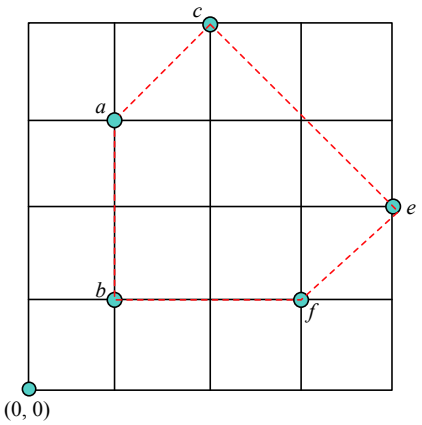

图 3　样本 S 背景基和背景分布

Fig. 3　Sample S background base and background distribution

结合算法和算例, 总结 IBBE 算法的优点:

①IBBE 算法极大简化了计算难度和编程难度;

②算法初始化阶段不需要依赖各因素极值;

③极大降低了基点数量, 有利于高维度数据提取基点。

3　基点分类算法

人工智能对数据的分类其实质是对同一类别的数据形成概念。本文提出的基点分类算法 (base point classification algorithm), 简称 BPC 算法, 以背景基提取算法为核心。通过对不同类别样本进行分类打包和学习, 最终将给定样本归纳成以各类别为单位的认知包, 一个认知包就是一个概念。

3.1　基本定义

BPC 算法训练得到的模型相对固定。在实际中, 决策者的主观因素常常扮演重要角色, 基于此, BPC 算法在做决策时需要加入决策者主观参数, 由主观参数 λ 变换得到的背景基称为 λ-背景基。

定义 7　设集合 $B_i\neq\varnothing$, 由 c 个类别生成的背景基 $\boldsymbol{B}=\bigcup_{i=1}^{c}\boldsymbol{B}_i$, 各类别用 i 表示, $i=1,2,\cdots,c$, $b_j\in\boldsymbol{B}$, 则有

$$[\boldsymbol{B}]_i=\{\boldsymbol{b}_j^i|i\in c,j=1,2,\cdots,n\},\tag{4}$$

其中 $n=|[\boldsymbol{B}]_i|$, 称 $[\boldsymbol{B}]_i$ 为类别 i 的背景基。

命题 1　如果 $[\boldsymbol{B}]_i$ 为类别 i 的背景基, 则 $\boldsymbol{B}_i=[\boldsymbol{B}]_i$。

证明　设 $[\boldsymbol{B}]_i$ 为类别 i 的背景基, $i=1,2,\cdots,c$, $\boldsymbol{B}$ 为由 c 个类别生成的背景基, 则有

$$\boldsymbol{B}=\bigcup_{i=1}^{c}[\boldsymbol{B}]_i$$

而由定义 7 得

$$\boldsymbol{B}=\bigcup_{i=1}^{c}\boldsymbol{B}_i$$

因此有

$$\boldsymbol{B}_i=[\boldsymbol{B}]_i$$

证毕。

定义 8　设 $\boldsymbol{B}$ 为 c 个类别的背景基, $[\boldsymbol{B}]_i$ 为类别 i 的背景基, $\boldsymbol{b}_j\in[\boldsymbol{B}]_i$, 向量 $\boldsymbol{\Lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_c)$, $\lambda_i>0$, $i=1,2,\cdots,c$, 使得

$$[\boldsymbol{B}_\lambda]_i=\boldsymbol{\Lambda}\otimes[\boldsymbol{B}]_i=\{\lambda_i\boldsymbol{b}_j|j=1,2,\cdots,|[\boldsymbol{B}]_i|\}\tag{5}$$

则称 $\boldsymbol{B}_\lambda=\bigcup_{i=1}^{n}[\boldsymbol{B}_\lambda]_i$ 为 λ-背景基。

从几何学角度看, λ-背景基 $\boldsymbol{B}_\lambda$ 是普通背景基 $\boldsymbol{B}$ 的放大或缩小。当 $\lambda\in[0,1)$ 时, 背景基 $\boldsymbol{B}$ 锁定的区域 $R(\boldsymbol{B})$ 包含 $\boldsymbol{B}_\lambda$ 锁定的区域 $R(\boldsymbol{B}_\lambda)$, 即 $R(\boldsymbol{B}_\lambda)\subset R(\boldsymbol{B})$; 当 $\lambda=1$ 时, 满足关系 $R(\boldsymbol{B}_\lambda)=R(\boldsymbol{B})$; 当 $\lambda\in(1,\infty)$ 时, $R(\boldsymbol{B}_\lambda)\supset R(\boldsymbol{B})$。综上所言, 对于区域 R 关系满足:

$$R=\begin{cases}R(\boldsymbol{B}_\lambda)\subset R(\boldsymbol{B}), & \lambda\in[0,1)\\ R(\boldsymbol{B}_\lambda)=R(\boldsymbol{B}), & \lambda=1\\ R(\boldsymbol{B}_\lambda)\supset R(\boldsymbol{B}), & \lambda\in[1,\infty]\end{cases}$$

实现 λ-背景基具体方法见定义 9。

定义 9　设 O 为第 i 类背景基 $[\boldsymbol{B}]_i$ 的重心, 其中 $m=|[\boldsymbol{B}]_i|$, $j\in m$, $\boldsymbol{b}_{ij}\in[\boldsymbol{B}]_i$, 则

$$\boldsymbol{b}_{ij}^{\lambda}=O+\lambda(O-\boldsymbol{b}_{ij})\tag{6}$$

这里 $[\boldsymbol{B}_\lambda]_i=\bigcup_{j=1}^{m}\boldsymbol{b}_{ij}^{\lambda}$。

3.2　构造基点分类算法

据第 2 章的 IBBE 算法和第 3.1 节中相关定

义, 本节设计一种新型的分类算法: 基点分类算法 (BPC 算法)。下面采用图示方法对 BPC 算法的构造原理进行详细阐述和说明。

如图 4 所示: 图 (a) 为待分类的数据集, 不同符号代表不同的类别, 共 3 个类别; 图 (b) 呈现了由 IBBE 算法提取的每个类别的基点; 为了形象表示基点围成的区域, 图 (c) 表示在图 (b) 基础上围成凸多边形, 其中该凸多边形可用来近似代替各类别的背景分布, 其中"☆"表示每个类别的重心 (由样本生成, 而非由基点生成), 这在一定程度上还原了样本背景分布的情形; 图 (d) 表示的是 λ-背景基的求解过程, 其中实线段围成的凸多边形是 λ-背景基; 图 (e) 就是分类器最后呈现出的背景基, 由 $0<\lambda<1$ 和 $\lambda=1$ 两层背景基构成, 可以形象区别为边缘区域和核心区域。

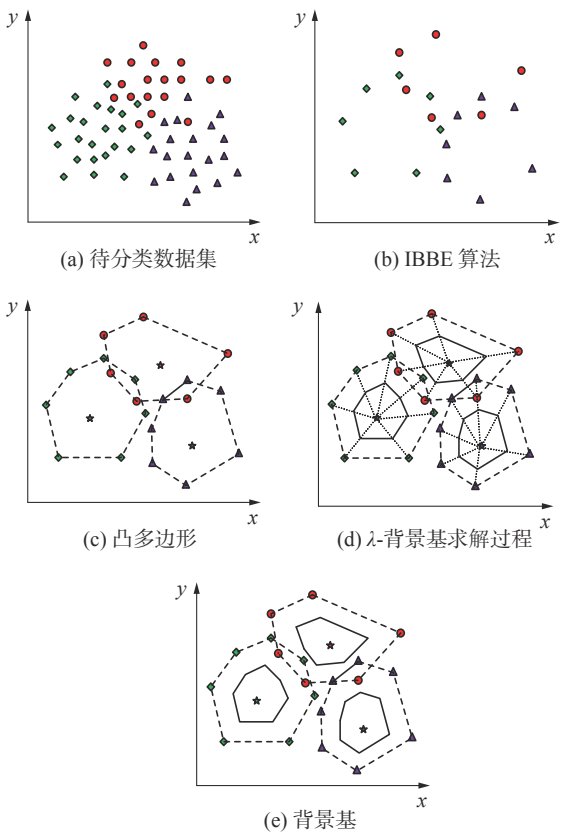

图 4　基点分类原理

Fig. 4　Base point classification principle

基于以上思想, 下面给出 BPC 算法的具体步骤。

算法　基点分类算法 (BPC 算法)。

输入　c 个类别的样本集合 S, 各类别用 i 表示; 参数向量 $\boldsymbol{\Lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_c)$, $\lambda_i>0$, $i=1,2,\cdots,c$。

初始化　背景基 $\boldsymbol{B}=\varnothing$。

1) 从 S 中取样本 x_k, 判定 x_k 类别 i; 从 $\boldsymbol{B}$ 中提取类别为 i 的背景基 $[\boldsymbol{B}]_i$。

2) 利用函数 IBBE(x_k, $[\boldsymbol{B}]_i$), 更新 $\boldsymbol{B}$ 中原来的背景基 $[\boldsymbol{B}]_i$, 具体做法为

$$[\boldsymbol{B}]_i' = \mathrm{IBBE}(x_k,[\boldsymbol{B}]_i)$$

$$\boldsymbol{B} = \mathrm{update}([\boldsymbol{B}]_i',[\boldsymbol{B}]_i)$$

函数 update() 功能是更新 $\boldsymbol{B}$ 中 $[\boldsymbol{B}]_i$。

3) 计算每个类别的重心 O。

4) 对 $\boldsymbol{B}$ 中每个类别做 λ 变换得到 $[\boldsymbol{B}_\lambda]_i$。

5) 输出 $\boldsymbol{B}_\lambda$。

设 M 表示样本个数, N 表示因素个数, p 表示基点个数。当样本数量较少时, p^2 与 M 的数量级接近; 当样本数量远远大于 N 和 p^2 时, 此时 N 和 p^2 可以忽略不计, 时间复杂度为 $O(MNp^2)\approx O(M)$。因此, BPC 算法在处理大样本数据时优势明显。

4　数值实验

采用两个实验对 BPC 算法进行测试。实验 1 采用二维数据作测试样本, 用来说明 BPC 算法在对二维数据分类时效果良好, 且有发现新类别的优点。实验 2 用三维数据作测试样本, 用来说明 BPC 算法在对三维数据分类效果同样良好, 并突出了 λ–背景基的在预测时的优势。

4.1　二维数据

为了验证 BPC 算法的可行性以及实际分类效果, 实验 1 采用二维数据作为样本数据。数据由 Matlab 软件仿真得到, 仿真数据为均匀分布, 每类样本集均为凸集, 有 2 个因素, 3 个类别, 每个类别 100 个样本, 共计 300 个样本。不同类别用不同符号标记, 如图 5 的图例所示。

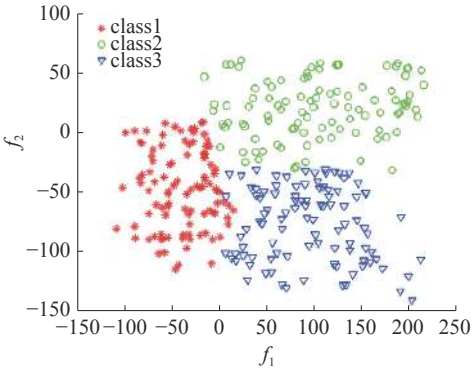

图 5　样本分布

Fig. 5　Sample distribution

图 5 为 3 个类别样本集的分布情况; 图 6 展现由基点分类法找到的各类别样本的基点, 3 个类的基点用不同的符号表示; 图 7 是删除各类别内点保留基点的示意图。

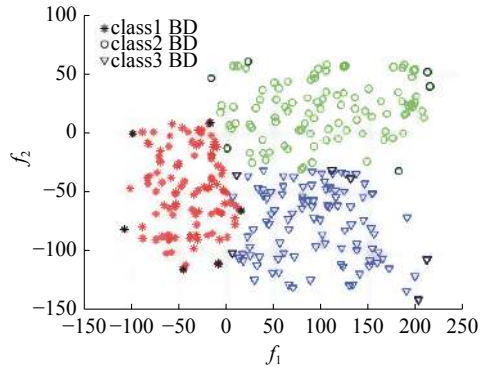

图 6　二维基点分布

Fig. 6　2D base point distribution

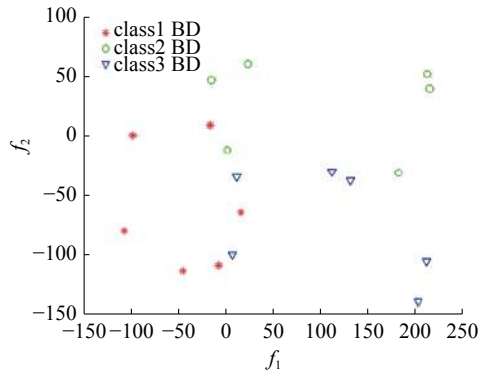

图 7　提取样本基点

Fig. 7　Get the sample base point

BPC 算法采用基点限定区域原理学习和预测，由于其原理与 SVM 算法原理具有相似性，所以实验结果与 SVM(libsvm) 算法进行比较是合理的。训练数据与测试数据为同一组数据。通过 Matlab 编程得到 BPC 算法的训练函数和预测函数。实验结果表明，BPC 算法准确率为 98.33%，SVM 算法准确率为 100.00%，两者结果相当。从实验结果来看，BPC 算法在开始阶段表现良好。

除此之外，通过分析实验，还能得到以下结论：

1) 在线性可分的样本分布下，BPC 算法与 SVM 算法预测结果相近，但 BPC 算法原理更简单易懂、实验结果可解释性强，在一定程度上更符合人的思维逻辑。

2)BPC 算法分类是通过构建每个类别的背景基来限定各类别区域范围，这种划分原理比 SVM 算法用超平面划分原理更加精准。当出现新类别样本时，BPC 算法可以及时发现远离它的类别，并为其定义新标签，而 SVM 算法只会将该样本归类到已学习到的类别中，不会产生新类别。所以，BPC 算法的一个非常明显的优势是在预测中可以发现新类别。

4.2　三维数据

实验 2 数据由 Matlab 仿真结果得到。仿真数据特点：每类数据集均为凸集，3 个因素，3 个类别 class1、class2 和 class3，每一类 220 个样本，其中 class1 与 class2 数据为均匀分布，class3 为非均匀分布，3 个类别线性可分，见表 1。

表 1　实验 2 数据说明

Table 1　Data description of experiment 2

类别	数据分布	因素个数	各类别样本数量	数据特点
class1	均匀分布	3	222	凸集
class2	均匀分布	3	222	凸集
class3	非均匀分布	3	222	凸集

图 8 展示了数据集的分布情况，不同符号代表不同类别。

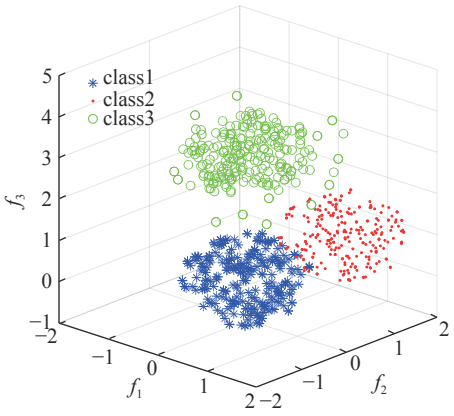

图 8　样本分布视角 1

Fig. 8　Samples distribution

BPC 算法可以提取三维数据的背景基。图 9 展示的是用 BPC 算法提取到实验 2 数据的背景基。由于 BPC 算法继承了 IBBE 算法的优点，所以提取到的基点数量少，却又能高效覆盖所要表达的信息。从图 10 可以看出，各类别的背景基数量远小于样本边界的数量。

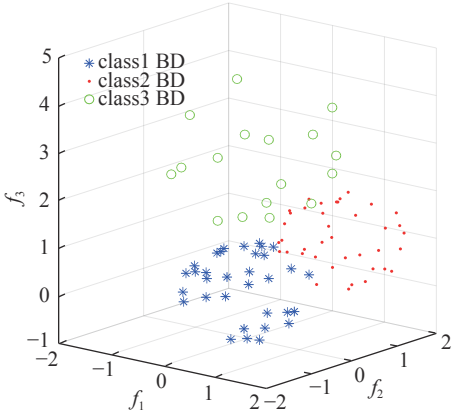

图 9　提取样本基点

Fig. 9　Get the sample base point

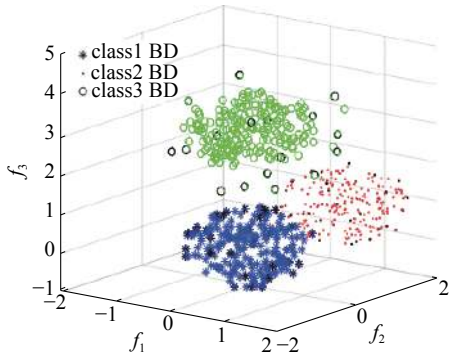

图 10　三维样本基点分布

Fig. 10　3D sample base point distribution

通过学习和预测实验 2 的数据, 实验得出: BPC 算法准确率为 97.73%, SVM 算法准确率为 99.09%。从结果来看, BPC 算法结果与 SVM 算法结果相当。但是从预测模型来看, BPC 算法比 SVM 算法更具有灵活性, 这主要体现在 λ-背景基的优势。

从图 8 可以看出, class3 样本集的外围数据明显处于游离状态, 其结合紧密程度远不如核心区域, 采用 λ-背景基, 可以缩小区域范围, 使其决策的可信度更高。

通过试错实验方法, 最后得到当 λ=0.7 时背景基能刚好包含核心数据。图 11 为 class 3 数据集的背景基 $\boldsymbol{B}$ 和 λ-背景基 $\boldsymbol{B}_\lambda$ 的示意图, 此时 $0<\lambda<1$, $\boldsymbol{B}_\lambda \subset \boldsymbol{B}$。通过调节$\lambda$参数, 改变由基点限定的区域, 这有利于抓取数据集核心特征。该方法除了分类应用以外, 也可为决策者提供决策依据。

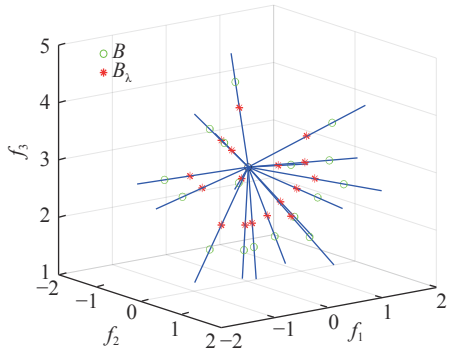

图 11　class3 的背景基 $\boldsymbol{B}$ 与 λ-背景基 $\boldsymbol{B}_\lambda$

Fig. 11　Background base $\boldsymbol{B}$ and λ-background base $\boldsymbol{B}_\lambda$ of class3

4.3　总结

BPC 算法的特点如下:

1) 上述实验只是针对二维、三维数据进行的, 实际上, BPC 算法可以扩展到 n 维数据集上, 且时间复杂度随着维度增加呈线性增长, 避免了高维数据基点冗余的现象。

2) BPC 算法具有智能的生长机制。由于 BPC 算法可以发现新类别, 新类别可以衍生新概念。从机制主义人工智能理论出发, 机器能产生新概念则具有自我学习和扩展学习的能力。据此可得, BPC 算法是具有一定智能生长机制的。

3) BPC 算法使用 λ-背景基使得预测或决策更灵活。分类器重复训练模型的代价很大, 为了减小代价、充分利用已训练模型, λ-背景基可以在原训练模型基础上, 根据需要扩大或缩小预测区域, 这种处理可以减少重复训练的代价, 同时可以继承原有知识。

4) BPC 算法原理完全由智能数学描述, 其算法和原理较传统数学更加简单高效。

5) BPC 算法得到的模型不是“黑箱”, 而是具有可解释的、符合人类逻辑思维的“白箱”。理论上, 该模型是决策者思维的同构映射。

5　结束语

本文从因素空间的背景基理论出发, 首先改进了背景基提取算法, 简化了计算步骤, 降低了时间复杂度, 从整体上提高了算法效率。其次, 构造了基点分类算法, 并通过 2 个实验展示了该算法在线性可分数据集的独特优势。下一步工作将研究非线性数据集处理、非凸数据集类别处理和精准识别新类别等。主要工作思路有: 1) 对于线性不可分数据集, 采用分块链式策略。将数据按照一定规则划分成块状, 然后用子背景基进行包络, 最终链式连接多个子背景基; 2) 对于非凸数据集类别, 可以采用因素映射方法, 即通过分析因素值变化的单调性进行相应的映射变换。

参考文献:

[1] 钟义信. 机制主义人工智能理论——一种通用的人工智能理论 [J]. 智能系统学报, 2018, 13(1): 2–18.
ZHONG Yixin. Mechanism-based artificial intelligence theory: a universal theory of artificial intelligence[J]. CAAI transactions on intelligent systems, 2018, 13(1): 2–18.

[2] 何华灿. 泛逻辑学理论——机制主义人工智能理论的逻辑基础 [J]. 智能系统学报, 2018, 13(1): 19–36.
HE Huacan. Universal logic theory: logical foundation of mechanism-based artificial intelligence theory[J]. CAAI transactions on intelligent systems, 2018, 13(1): 19–36.

[3] 汪培庄. 因素空间理论—机制主义人工智能理论的数学基础 [J]. 智能系统学报, 2018, 13(1): 37–54.
WANG Peizhuang. Factor space-mathematical basis of mechanism based artificial intelligence theory[J]. CAAI transactions on intelligent systems, 2018, 13(1): 37–54.

[4] 汪培庄. 因素空间与因素库 [J]. 辽宁工程技术大学学报(自然科学版), 2013, 32(10): 1297–1304.
WANG Peizhuang. Factor spaces and factor data-bases[J]. Journal of Liaoning Technical University (Natural Science), 2013, 32(10): 1297–1304.

[5] 汪培庄. 因素空间与数据科学 [J]. 辽宁工程技术大学学报(自然科学版), 2015, 34(2): 273–280.
WANG Peizhuang. Factor spaces and data science[J]. Journal of Liaoning Technical University (Natural Science), 2015, 34(2): 273–280.

[6] WANG Peizhuang, LIU Zengliang, SHI Yong, et al. Factor space, the theoretical base of data science[J]. Annals of data science, 2014, 1(2): 233–251.

[7] WANG Peizhuang, OUYANG He, ZHONG Yixin, et al. Cognition math based on factor space[J]. Annals of data science, 2016, 3(3): 281–303.

[8] 钟义信. 高等人工智能原理: 观念·方法·模型·理论 [M]. 北京: 科学出版社, 2014.

[9] 钟义信. 信息科学与技术导论 [M]. 3 版. 北京: 北京邮电大学出版社, 2015.

[10] 曲国华, 曾繁慧, 刘增良, 等. 因素空间中的背景分布与模糊背景关系 [J]. 模糊系统与数学, 2017, 31(6): 66–73.
QU Guohua, ZENG Fanhui, LIU Zengliang, et al. Background distribution and fuzzy background relation in factor spaces[J]. Fuzzy systems and mathematics, 2017, 31(6): 66–73.

[11] 曾繁慧, 郑莉. 因素分析法的样本培育 [J]. 辽宁工程技术大学学报(自然科学版), 2017, 36(3): 320–323.
ZENG Fanhui, ZHENG Li. Sample cultivation in factorial analysis[J]. Journal of Liaoning Technical University (Natural Science), 2017, 36(3): 320–323.

[12] 吕金辉, 刘海涛, 郭芳芳, 等. 因素空间背景基的信息压缩算法 [J]. 模糊系统与数学, 2017, 31(6): 82–86.
LV Jinhui, LIU Haitao, GUO Fangfang, et al. The algorithm of extraction of background bases[J]. Fuzzy systems and mathematics, 2017, 31(6): 82–86.

[13] 李建中, 李英姝. 大数据计算的复杂性理论与算法研究进展 [J]. 中国科学: 信息科学, 2016, 46(9): 1255–1275.
LI Jianzhong, LI Yingshu. Research progress in the complexity theory and algorithms of big-data computation[J]. SCIENTIA SINICA informationis, 2016, 46(9): 1255–1275.

[14] PAWLAK Z. Rough sets[J]. International journal of computer & information sciences, 1982, 11(5): 341–356.

作者简介：

蒲凌杰，硕士研究生，主要研究方向为因素空间理论、数据挖掘、智能决策。

曾繁慧，教授，主要研究方向为基于因素空间的数据挖掘理论与应用、模糊结构元理论与应用。参与完成中国工程院重点项目、国家自然科学基金项目、辽宁省基金项目、教育部高校博士学科点专项科研基金项目等。获多项省、市级奖励。发表学术论文 50 余篇。

汪培庄，教授，博士生导师，主要研究方向为模糊数学及其在人工智能中的应用，近期主要致力于因素空间在人工智能和数据科学中的应用研究。提出和创立了模糊集的随机落影表示、真值流推理和因素空间等数学理论，多次获得国家级和部委级奖励，获得一次国际奖。发表学术论文 200 余篇，出版学术著作 4 部。