文章编号：2095-2163(2023)06-0174-07 中图分类号：O29；C931.1 文献标志码：A

粒度信息决策算法之差转计算应用研究

赵 静[1]，包研科[2]

（1 黔南民族师范学院 数学与统计学院，贵州 都匀 558000；2 辽宁工程技术大学 理学院，辽宁 阜新 123000）

摘 要：分类决策问题是人工智能领域的核心问题，因素空间针对这一问题构建了相应的粒度信息决策算法，其中较为典型的有因素分析法及差转计算算法，这两个算法的本质是：根据单一条件信息与决策信息在论域中形成的等价类包含关系诱导出概念知识，但存在不能描述不同类型的因素在概念形成过程中的作用的局限。为解决这一问题，本文基于因素空间及商集合理论，定义了因素的析取、合取变换，在此基础上构造了数据粒度变换方法。为验证变换方法的有效性，以同为产生式推理算法的决策树算法为对比算法，在 UCI 经典数据集 Wisconsin Breast Cancer 上进行了实例验证，结果表明：数据粒度变换方法是有效的，知识挖掘形成经验推理系统的时间成本要低于决策树，经验推理系统泛化效果同决策树持平。

关键词：因素空间；决策算法；差转计算；知识挖掘；商集合

Research on the set subtraction and rotation calculation application of granular information decision making algorithm

ZHAO Jing[1], BAO Yanke[2]

(1 School of Mathematics and Statistics, Qiannan Normal University for Nationalites, Duyun Guizhou 558000, China;
2 College of Science, Liaoning Technical University, Fuxin Liaoning 123000, China)

【Abstract】 Classification decision-making issues are crucial to the field of artificial intelligence. Factor space constructs the corresponding granular information decision-making algorithm for this problem, among which the more typical algorithm is the factor analysis and the set subtraction and rotation calculation. The essence of these two algorithms is that conceptual knowledge is induced by the equivalent class inclusion relationship formed in the discourse domain between single condition information and decision information. However, there is a limitation that it is not possible to describe the role of different types of factors in the process of concept formation. In order to solve this problem, based on the theory of factor space and quotient set, this paper defines the analysis and combination transformation of factors, and constructs the data granularity transformation method. To verify the effectiveness of the transformation method, the decision tree algorithm, which is also a generative inference algorithm, is used as the comparison algorithm. Instance validation is performed on the UCI classic dataset Wisconsin Breast Cancer. The results show that the data granularity transformation method is effective, the time cost of knowledge mining to form an empirical reasoning system is lower than that of the decision tree, and the generalization effect of the empirical reasoning system is the same as that of the decision tree.

【Key words】 factor space; decision algorithms; set subtraction and rotation calculation; knowledge mining; quotient set

0 引 言

因素空间理论是汪培庄先生于 1982 年提出，旨在描述随机性和模糊性本质规律的数学理论，与认知科学交互，成为数据科学与智能科学的基础理论和概念与知识表达的普适性框架[1]。2014 年，汪培庄发起并主导了因素空间在数据科学中的应用问题的讨论。同年，包研科[2-3]在因素分析法的基础上，结合因素空间理论，提出名为“差转计算（The set subtraction and rotation calculation，S&R）”的多因素决策算法，并深入讨论了其决策过程的深层理论背景和几何结构。

差转计算以有监督类数据为操作对象，挖掘其内部蕴含的规律，表现为“*if*…，*then*….”结构的规

基金项目：贵州省教育厅高等学校科学研究项目（黔教技[2022]378 号）。

作者简介：赵 静（1996-），男，硕士，讲师，主要研究方向：因素空间理论下知识发现理论与应用；包研科（1962-），男，硕士，副教授，主要研究方向：工程数据分析、统计预测与决策、统计机器学习等。

通讯作者：包研科 Email：baoyanke-9257@163.com

收稿日期：2022-06-16

则模型，其决策机制是建立在人脑解决分类问题的认知原理之上，在定性因素的多因素分类问题中取得了较好的实测效果[2]。2017 年，包研科[4]给出定量因素上的差转计算修正算法，拓展了差转计算的应用场景。自差转计算提出以来，笔者所在团队进行了大量的算法测试与验证工作。文献[4-5]以及非公开的算法测试结果表明，差转计算算法在定性变量的样本数据集上，在规则挖掘效率和泛化性能上优于定量变量的样本数据集。

因素空间理论认为决策问题是问题研究所在论域内概念的认知过程[5]。决策过程实质是论域内研究对象的分类识别问题，从人工认知学角度来讲，决策问题应当在条件信息数据与决策信息数据支撑下回到论域中进行讨论样本的归属。基于此，差转计算的决策机制是构建条件信息数据与决策信息数据到论域的凸包，并基于凸包内样本数据建立决策准则，诱导出“*if*…，*then*…”语句的推理知识，同决策树类似。差转计算的基本决策过程是基于统计信息构建数据空间到认知本体的划分类，对于差转计算的某一次决策过程，划分形成等价类有 3 类关系如图 1 所示。

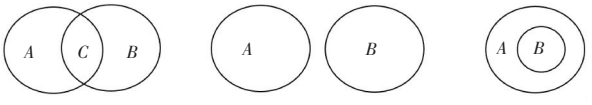

图 1　等价类的 3 种关系

Fig. 1　Three relationships of equivalence classes

图 1 中 A、B、C 表示差转计算依条件信息数据与决策信息数据对论域形成的划分类。图 1(b)、(c)两类关系可归结为线性可分和完全包含关系的分类识别问题，现有算法如感知机、支持向量机等能够很好解决此类问题。关键问题在于如何有效识别图 1(a)中 C 区域中对象的归属。

不同分类算法在处理区域 C 时采用的方法不同，应用较为广泛的方式是基于样本数据采用降维技术将数据变换到区域 A 或 B 中，存在的风险是数据失真，进而可能导致分类结果不准确的情况。在保证数据不失真的情况下，还能有效识别区域 C 内对象归属问题，可以考虑采用升维的方法，从因素空间理论的角度来讲，需要引入新的指标，即隐性条件因素的发现，文献[5]在这个方面做了突破性工作，但尚未进行深入研究。

本文基于因素空间和商集合理论背景下，提出一种基于信息粒度变换的数据处理方法，本质是变换数据观测视角，在不同视角下对论域进行划分，形成有效决策知识。为验证本文所提方法的有效性，将该方法结合差转计算，并应用于样本数据为离散型的多因素分类决策问题中。

1　基本概念及差转计算算法原理

分类问题是人脑认知过程的主体，分类算法的设计过程应符合人工认知的 4 个基本原理[5]。为方便理解本文工作，就差转计算所涉及概念进行阐述。

1.1　因素及其运算

科学研究是在特定研究范围内对存在的问题进行证伪的过程，根本目的在于形成符合人类认知结构的知识，而知识由概念表达，概念的阐述由内涵和外延对合构成。差转计算的根本目的是基于统计信息挖掘内在规律性知识。

定义 1　称问题研究过程中讨论的所有对象 u_i 构成的可列集合 U 为论域，记为式(1)：

$$U = \{u_i\}_{i=1}^{\infty} \tag{1}$$

统计研究过程中，论域 U 内研究对象是有限的，即式(2)

$$U = \{u_i\}_{i=1}^{n}, n \in N^+ \tag{2}$$

定义 2　称满映射，即 $f: U \to I_f$ 为定义在 U 上的因素，即对 $\forall u_i \in U$，在 I_f 存在一个性态特征 d，使得 $f(u_i) = d$，I_f 是论域 U 中对象 u_i 在 f 上表达出的性态特征构成的集合，称为 f 的相态空间。

实际应用过程中，性态 d 存在空置的风险，即统计样本存在缺失值。因素空间理论认为缺失值亦是特殊相态值。

因素按照其发挥的功能可分为条件因素和结果因素。本文中若无特殊指代，一般因素指的是条件因素。以有监督数据为例，类别或标签数据对应指标即为结果因素，提供决策信息；除结果因素以外的指标称为条件因素，提供决策条件参考信息。按照相态空间数据属性可分为连续型因素和离散型因素。

定义 3　设映射 $\overleftarrow{f}: I_f \to 2^U$，满足：$\forall d \in I_f$，$\overleftarrow{f}(d) = [d]_f \in U/f \subset 2^U$，称 $\overleftarrow{f}$ 为因素 f 的回溯(Recall)。

回溯是因素的广义逆映射。回溯概念的定义为条件信息数据与决策信息数据回到论域中进行构建划分提供了方法和工具。因此，由因素及回溯定义，不论是条件因素，还是结果因素，均具有两个功能：

(1)因素是概念的本位属性限定工具，即式(3)和式(4)：

$$[d]_f = \{u_i \mid f(u_i) = d \in I_f\} \subseteq U, i = 1,2,\cdots,n \tag{3}$$

$$f([d]_f) = d \tag{4}$$

其中，$[d]_f$ 是对象 u_i 由因素 f 的回溯 $\overleftarrow{f}$ 在论域 U 中构成的等价类，是概念外延在数据科学中的表达。

（2）因素是论域的划分工具，设 $I_f = \{d_1, d_2, \cdots, d_s\}$，则商集 $U/f = \{[d_1]_f, [d_2]_f, \cdots, [d_s]_f\}$ 是 U 的一个划分，构成概念外延集合。显然式(5)和式(6)成立。

$$\forall\ [d_j]_f \in U/f, j = 1,2,\cdots,s, \bigcup_{j=1}^{s} [d_j]_f = U \tag{5}$$

$$\forall\ [d_j]_f, [d_k]_f \in U/f, [d_j]_f \cap [d_k]_f = \varnothing \tag{6}$$

上述定义不仅建立了基于因素空间理论的数据挖掘算法独特的数学基础框架，亦可保证基于数据的概念表达内涵描述与外延描述的一致性。

1.2 算法原理及步骤

差转计算的算法本质就是将决策信息和条件参考信息构成的本体关系映射到条件因素上。

定义 4 设论域 U 中样本容量为 n，称对各个样本 u_i 的顺序编号构成的集合 $K = \{1,2,\cdots,i,\cdots,n\}$ 为 U 的秩序集。若 A 是 U/f 中任意的含有 s 个对象的等价类，记 $K^{(A)} = \{i_1, i_2, \cdots, i_s\} \subseteq K$ 为 A 的秩序子集，则称 $(K^{(A)}, f(A))$ 为 A 的 f — 表征，记为 $R_f(A)$。

定义 5 设 f 和 g 是同时定义在论域 U 上的因素，对 $\forall l \in I_g$，称 $R_f([l]_g)$ 为等价类 $[l]_g$ 在 f 上的踪影(Representation)。

定义 6 称 $[l]^{(k)}_{g|sortf} \subset [l]_g$ 为因 f 的相态排序而形成的第 k 个聚集子块(简称子块)，满足式(7)和式(8)：

$$[l]^{(k)}_{g|sortf} \cap [l]^{(r)}_{g|sortf} = \varnothing, \bigcup_{\forall k} [l]^{(k)}_{g|sortf} = [l]_g \tag{7}$$

$$\forall p,q \in I_g, p \neq q, [p]^{(k)}_{g|sortf} \cap [q]_g = \varnothing \tag{8}$$

定义 7 设 f 和 g 是同时定义在论域 U 上的因素，其相态空间分别记为 I_f、I_g，且 I_f 存在序关系，若 $\exists t \in I_f, \exists l \in I_g$，满足式(9)：

$$R_f([t]_f) \subseteq R_f([l]^{(k)}_{g|sortf}) \tag{9}$$

则称式(10)

$$\alpha_{f\to g} = pro\Big(\sum_{\forall t \in I_f} (R_f([t]_f) \subseteq R_f([l]^{(k)}_{g|sortf}))\Big) \tag{10}$$

为因素 f 对 g 的预测系数或决定度，并称 $[t]_f$ 是 $[l]_g$ 的 f 决定类，其中 $pro(\bigstar)$ 表示对 $\bigstar$ 求统计频率。若 $\alpha_{f^*\to g} = \max_{\forall f} \alpha_{f\to g}$，则称 f^* 为优势因素。

为清楚阐述定义 4～定义 7 描述概念及其转换关系，假定对因素 f 的相态排序后形成如图 2 所示的概念转换示意图。

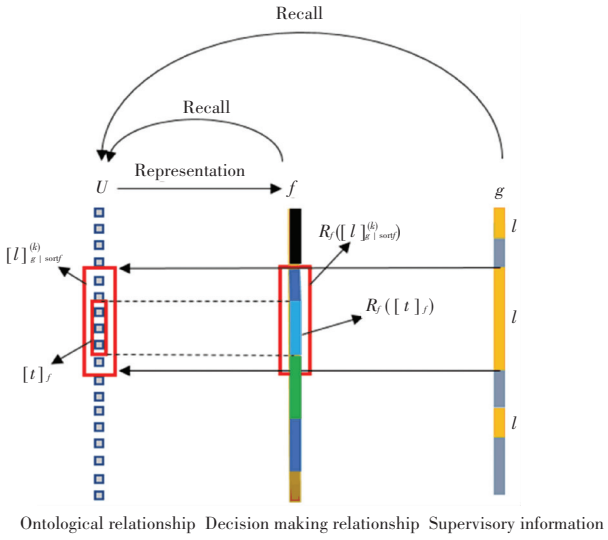

图 2 概念转换示意图

Fig. 2 Concept conversion schematic

图 2 中，$[l]^{(k)}_{g|sortf}$ 表示等价类 $[l]_g$ 由于因素 f 的相态空间排序而形成的第 k 个子块。决策问题是关于论域中对象归属问题的讨论，因素 g 表征了决策信息，只能提供决策参考，从因素回归到论域中进行讨论是因素空间的基本思想。因此，决策的本体关系是 $[t]_f \subseteq [l]^{(k)}_{g|sortf} \subseteq [l]_g$，但决策的观察是发生在因素 f 上的，算法的表述是 $R_f(t) \subseteq R_f([l]^{(k)}_{g|sortf})$，差转计算算法原理如下：

差转计算的算法原理[2]假定 f 和 g 是论域 U 上的离散型因素，由 f 和 g 定义的两个概念记为

$\varepsilon_f(u_i,k) = (k, [k]_f)$ 和 $\varepsilon_g(u_i,r) = (r, [r]_g)$，则复合推理句 $\forall\ u_i \in U$，若 $f(u_i) = k$，则 $g(u_i) = r$，恒真的充分必要条件是 $[k]_f \subset [r]_g$。

基于前述定义及原理，差转计算实现了由样本数据空间到研究论域的分类决策过程，差转计算在分类决策过程中的算法步骤：

算法 差转计算

输入 有限样本数据集 S 见表 1。

初始化：复合推理知识集合 $R = \varnothing$；样本集内样本个数 $|S|$；

过程：差转计算算法

(1)判断 $|S|$ 是否大于 0。是，计算每个因素 $f_i, i = 1,2,\cdots,m$ 的决定度；否，算法结束；

(2)定位优势因素 f^*，记其决定度为 δ_{f^*}，转(3)；

(3)判断 δ_{f^*} 是否大于 0。否，则转(5)，算法结束；是，依据 $R_{f^*}([t]_{f^*}) \subseteq R_{f^*}([l]^{(k)}_{g|sortf^*})$ 辨识决定类，则转(4)；

表 1　有限样本数据集 S

Tab. 1　Limited sample data sets

U	f_1	f_2	…	f_i	…	f_m	g
u_1	$f_1(u_1)$	$f_2(u_1)$	…	$f_i(u_1)$	…	$f_m(u_1)$	$g(u_1)$
u_2	$f_1(u_2)$	$f_2(u_2)$	…	$f_i(u_2)$	…	$f_m(u_2)$	$g(u_2)$
⋮	⋮	⋮		⋮		⋮	⋮
u_j	$f_1(u_j)$	$f_2(u_j)$	…	$f_i(u_j)$	…	$f_m(u_j)$	$g(u_j)$
⋮	⋮	⋮		⋮		⋮	⋮
u_n	$f_1(u_n)$	$f_2(u_n)$	…	$f_i(u_n)$	…	$f_m(u_n)$	$g(u_n)$

注：表中 f_i 代表第 i 个条件因素，g 代表结果因素。

(4)形成推理语句 R'：若 $f^*(u_i)=t$，则 $g=l$，将 R' 加入到 R 中，在 S 中删除 $[t]_{f^*}$ 对应数据，更新 S，则转(1)。

输出　复合推理知识集合 R，算法结束。

由上述过程可以看出，差转计算算法结束的要求有：

(1)删空操作数据集；

(2)在数据集非空情形下不能形成有效决定类，此类情形极容易造成决策信息丢失，进而影响算法泛化效果，本文核心工作在于利用粒度信息变换去解决此类情形下的决策问题。

另外，因素空间中因素分析法与差转计算法在规则知识发现的过程中区别是第(4)步骤的不同，因素分析法既要删除因素分析表中对应因素，又要删除因素分析表中决定类对应数据，而差转计算仅删除决定类对应数据。从认知学角度来讲，差转计算更符合人脑辨识事物的过程。

2　粒度变换的思想和原理

分类决策研究最核心的问题在于辨识图 1 中区域 C 内的对象的归属，存在不能描述不同类型的因素在概念形成过程中的作用的局限[5]。

因素空间中对于区域 C 的对象归属一般采用不同视角进行辨识，即利用不同因素构建等价类包含关系，但亦避免不了上述局限性。数据科学中，大多算法对区域 C 的辨识，在实际应用中一般采用指标间相关性进行降维，将数据变换到区域 A 或 B 中，从而实现有效分类，但存在数据失真的问题，导致最终分类结果不准确的风险；值得一提的是决策树算法在此类问题中表现出较好的特性，其决策过程依赖指标同样本之间的统计数据构造指标分类节点，但亦存在依靠“贪心策略”得出的分类节点不可回溯，某些树枝对应的规则的支持度相对较小，从而规则的可靠性也较小，很有可能使推理出来的规则存在系统误差的问题[4]；另外一种依赖统计数据构造推理知识的算法是关联规则，其基本过程是采用水平组织逐层搜索的迭代方法形成推理知识，但其过度依赖数据样本的背景空间，推理出的知识亦仅局限于数值关系表达，不符合人脑辨识事务的基本过程[6]。

为此，本文构造基于统计数据中因素升维观测的粒度变换方法，进而结合差转计算实现多因素知识挖掘过程。

2.1　粒度变换原理

因素空间中维度变换由因素的析取与合取运算实现。

定义 8(因素的析取)　设 f 和 g 均为定义在论域 U 上的因素，称运算 $h=f\wedge g=g\wedge f$ 为因素 f 和 g 的析取运算。

因素的析运算旨在描述因素 f 和 g 在概念形成过程中的协同作用，进一步可理解为概念的分化功能，即析取因素越多，划分的概念越明确。

其概念外延划分功能由下述定理实现：

定理 1[5](析运算基本定理)　$U/f\wedge g\Leftrightarrow U/f\circ U/g$，其中 $U/f\circ U/g$ 为商集 U/f 与 U/g 的积。

定义 9(因素的合取)　设 f 和 g 均为定义在论域 U 上的因素，称运算 $c=f\vee g=g\vee f$ 为因素 f 和 g 的合取运算。

因素的合取描述了因素 f 和 g 在概念形成过程中的协同作用，进一步可理解为概念的同化功能，即因素析运算中因素越多，划分的概念越“粗糙”。

概念外延划分功能由合运算基本定理实现。

定理 2[5](合运算基本定理)　$U/f\vee g\Leftrightarrow U/f+U/g$，其中 $U/f+U/g$ 为商集 U/f 与 U/g 的和。

上述定义及定理既实现了因素空间到数据空间的联立，又实现了概念内涵和外延之间的对合关系，图 1 中区域 C 内对象的辨识问题实质是分化概念对其外延的对合描述问题，是分化和同化的暂时平衡。

定义 8、定义 9 和定理 1、定理 2 为新因素的引入及其相态空间的构造提供了方法,新因素的引入旨在变换观测角度,或提升观测维度。

2.2 粒度变换步骤

差转计算在知识挖掘过程中以决定度为决策准则,依据人工认知基本原理,决定度过低不利于推理知识的形成,决定度过高则影响泛化性能;决定度过低,其样本反变到图 1 中区域 C 内,无法进行有效辨识,此时,由定义 8 可联立因素进行升维观测,旨在发现能够辨识样本的知识,升维的过程既涉及新因素的引入,又涉及新因素相态空间的构造,实质是利用原始样本数据重构数据空间。

为描述简单,假设数据升维观察仅涉及两两因素联立,多因素联立过程以以下步骤外推得到。设 D 是论域 U 上原始有限样本数据集, $\forall f_i, f_j, i \neq j$ 为定义在 D 上的离散型条件因素,其相态空间分别为 I_{f_i}, I_{f_j} , g 为定义在 D 上的结果因素。不失一般性,记 $I_g = \{1,2,\cdots,s\}$,重构样本数据空间过程中结果因素及其相态空间保持不变。重构步骤如下:

输入 有限样本数据集 D

过程:数据升维算法

步骤 1 对因素 $f_i, f_j, i \neq j, \forall x \in I_{f_i}, y \in I_{f_j}$, 分别求因素 f_i, f_j 在 D 内的等价类,即 $A = [x]_{f_i} = \{A_1, A_2, \cdots, A_n\}$, $B = [y]_{f_j} = \{B_1, B_2, \cdots, B_l\}$,n、l 分别代表I_{f_i}, I_{f_j} 中相态种类数;

步骤 2 求集族 $C = \{C_{p,q} \mid C_{p,q} = A_p \cap B_q, A_p \in A, B_q \in B, p = 1,2,\cdots,n, q = 1,2,\cdots,l, A_p \cap B_q \neq \varphi\}$;

步骤 3 求不相交并集族 $\pi(C)$,即若 $\exists C_z, C_r \in C, C_z \neq C_r, C_z \cap C_r \neq \phi$, 在 C 中删除 C_z, C_r 并将 $C_z \cup C_r$ 加入到 $\pi(C)$ 中;

步骤 4 记联立因素为 $h_{i,j}$, 此步骤为构造相态空间 $I_{h_{i,j}}$。由步骤 3,记 $\pi(C) = \{C_1, C_2, \cdots, C_s\}$,则 $\forall u \in C_a, a = 1,2,\cdots,s, f_i \wedge f_j(u) = a$, 即有 $I_{h_{i,j}} = \{1,2,\cdots,s\}$。

为方便理解重构过程,以实例结合进行阐述。

记待挖掘数据集 D 见表 2。

显然,差转计算决策准则此时不适用,无法进行有效知识挖掘,需进行数据升维观测,旨在联立因素扩大解析能力,形成有效决定类。

输入 有限样本数据集 D, 即表 2

步骤 1 因素 f_1 在论域中形成的等价类为 $A = \{\{u_1, u_2\}, \{u_3, u_4\}, \{u_5, u_6\}\}$;因素 f_2 在论域中形成的等价类为 $B = \{\{u_4, u_5\}, \{u_1, u_3\}, \{u_2, u_6\}\}$;

步骤 2 求集族,记 $A_1 = \{u_1, u_2\}$, $A_2 = \{u_3, u_4\}$, $A_3 = \{u_5, u_6\}$, $B_1 = \{u_4, u_5\}$, $B_2 = \{u_1, u_3\}$,$B_3 = \{u_2, u_6\}$

表 2 待挖掘数据集 D

Tab. 2 The dataset D to be mined

	f_1	f_2	g
u_1	1	2	1
u_2	1	3	2
u_3	2	2	2
u_4	2	1	3
u_5	3	1	1
u_6	3	3	3

则有:

$A_1 \cap B_1 = \phi, C_1 = A_1 \cap B_2 = \{u_1\}$;$C_2 = A_1 \cap B_3 = \{u_2\}$

$C_3 = A_2 \cap B_1 = \{u_4\}$,$C_4 = A_2 \cap B_2 = \{u_3\}$,$A_2 \cap B_3 = \phi$

$C_5 = A_3 \cap B_1 = \{u_5\}$,$A_3 \cap B_2 = \phi, C_6 = A_3 \cap B_3 = \{u_6\}$

则 $C = \{C_1, C_2, C_3, C_4, C_5, C_6\} = \{\{u_1\}, \{u_2\}, \{u_4\}, \{u_3\}, \{u_5\}, \{u_6\}\}$;

步骤 3 求不相交并集族 $\pi(C)$,判断 C 内元素是否存在交集,依秩序求交并放入 $\pi(C)$ 内,显然 C 内元素不存在交集,则 $\pi(C) = C = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}\}$;

步骤 4 重构相态空间,记 $h_{1,2} = f_1 \wedge f_2$,则$I_{h_{1,2}}$ 为 $\{1,2,3,4,5,6\}$。

输出 变换数据集 D', 见表 3。

表 3 变换数据集 D'

Tab. 3 Transformed dataset D'

	$h_{1,2}$	g
u_1	1	1
u_2	2	2
u_3	3	2
u_4	4	3
u_5	5	1
u_6	6	3

显然,在数据集 D' 上可以进行差转计算,并且形成一一对应的知识类。

值得注意的是在较多因素下的因素升维变换观测很有可能存在组合爆炸的问题,特别是二因素组合带来的数据爆发式增长,极易带来模型应用复杂度的提升。针对这个问题,本文做出如下构想:可以基于因素轮廓构造决定灵敏综合信息度量,在此基础上可以衍生到在人工认知角度下的多因素适度组

合问题，能够很大程度上解决这一问题。

3　实证分析

为验证方法的有效性，将本文所提方法结合差转计算，同时以在知识挖掘过程同为推理算法的决策树作为对比算法，并以 UCI 共享数据库中较为经典的 Wisconsin Breast Cancer（简记为 WBC）为研究对象，采用留出法按比例 9∶1 将 WBC 数据集划分为训练集和测试集，分别利用两种算法对训练集进行知识挖掘，将挖掘出的知识应用于测试集，采用错误率、查准率、查全率、敏感度和 $F1$ 综合度量 5 个指标对最终结果进行比较。

本文的知识挖掘及模型泛化过程代码均基于 MATLAB 2016b 根据差转计算算法原理设计实现，对比算法决策树 C5.0 的知识挖掘和泛化过程在商用软件 see5 上完成。

3.1　数据说明

数据集 WBC 共计 699 个样本，包含 9 类条件因素和两类结果因素，条件因素名称、类型及值态分布见表 4。

表 4　数据集 WBC 条件因素名称及值态分布

Tab. 4　WBC's conditional factor name and value distribution

因素编号	条件因素英文名称	条件因素中文名称	因素类型	值态范围
(1)	Clump Thickness	肿块密度	离散型	{1,2,3,4,5,6,7,8,9,10}
(2)	Uniformity of Cell Size	细胞大小均匀性	离散型	{1,2,3,4,5,6,7,8,9,10}
(3)	Uniformity of Cell Shape	细胞形状均匀性	离散型	{1,2,3,4,5,6,7,8,9,10}
(4)	Marginal Adhesion	边界粘性	离散型	{1,2,3,4,5,6,7,8,9,10}
(5)	Single Epithelial Cell Size	单上皮细胞大小	离散型	{1,2,3,4,5,6,7,8,9,10}
(6)	Bare Nuclei	裸核	离散型	{1,2,3,4,5,6,7,8,9,10}
(7)	Bland Chromatin	微受激染色质	离散型	{1,2,3,4,5,6,7,8,9,10}
(8)	Normal Nucleoli	常态核仁	离散型	{1,2,3,4,5,6,7,8,9,10}
(9)	Mitoses	有丝分裂	离散型	{1,2,3,4,5,6,7,8,9,10}

数据集 WBC 有 1 个结果因素，包含 2 个相态良性（benign）和恶性（malignant），各有 458 和 241 个样例；共有 9 个条件因素，每个因素均有 10 个相态，相态值按细胞学特征从 1-10 分为 10 个等级，值态 1 最接近结果因素的相态 benign，值态 10 最接近结果因素的相态 malignant。

699 个样本以留出法按 9∶1 划分为训练集和测试集，在良性（benign）和恶性（malignant）两类别上的样本数据分布见表 5。

表 5　样本数据分布

Tab. 5　Sample data distribution

数据集	结果因素类别		合计
	良性（benign）	恶性（malignant）	
训练集	406	215	621
测试集	51	27	78
合计	457	242	699

3.2　测试结果

测试分为两个阶段，分别为知识挖掘阶段和知识泛化阶段。差转计算在训练集上知识挖掘经历两个过程，分别是单因素决策知识的提取和双因素析取运算下决策知识的挖掘，挖掘出的推理语句序列见表 6。

差转计算在训练集上共计操作 12 次，前 9 次为单因素规则知识挖掘，第 9 次知识挖掘后训练集非空，剩余 426 组样本数据；此时单因素不能形成有效决定类，需引入双因素进行信息粒度变换，变换后样本维度为 426×37，变换数据挖掘形成第 10~12 组双因素规则知识，这个过程仅 3 次便使差转计算收敛，收敛速度很快。整个知识挖掘过程共计执行 0.323 s，知识挖掘速度小于在 see5 上决策树知识挖掘时间（0.532 s）。

表 6 中知识的依次可解读为：若因素（2）取值为 5 或 10 时，则结果为（malignant）；若因素（2）取值不为 5 或 10 时，但因素（1）取值为 9 或 10 时，则结果为（malignant）。

从机器学习角度来讲，知识的泛化一般需要在训练集和测试集上分别进行。对于知识在训练集上的泛化，不论是从差转计算算法原理，亦或是笔者所作的大量实验来看，差转计算均能在训练集上实现高效决策，准确率几乎达到 100%，表明差转计算的学习过程不会产生冗余规则；但决策树并不能达到这一目标。

表 6 规则知识挖掘结果表

Tab. 6 The result of rule-knowledge mining

轮次	优势因素	决定性事件
1	(2)	5、10 → (malignant)
2	(1)	9,10 → (malignant)
3	(8)	9,10 → (malignant)
4	(6)	2 → (benign) or 6、9 → (malignant)
5	(7)	6、8、9、10 → (malignant)
6	(9)	5、8 → (benign) or 4、6 → (malignant)
7	(3)	9,10 → (malignant)
8	(4)	7,8 → (malignant)
9	(5)	7 → (benign)
10	(3) ∧ (6)	1 ∧ 1、3、4 → (benign) or 2 ∧ 1、3、5 → (benign) or 2 ∧ 4 → (malignant) or 3 ∧ 1、4 → (benign) or 3 ∧ 3、7、8 → (malignant) or 4 ∧ 7 → (benign) or 4 ∧ 4、10 → (malignant) or 5 ∧ 8、10 → (malignant) or 6 ∧ 8、10 → (malignant) or 6 ∧ 7、10 → (malignant) or 8 ∧ 4 → (benign) or 8 ∧ 4 → (malignant)
11	(1) ∧ (4)	1 ∧ 2 → (benign) or 1 ∧ 4 → (malignant) or 3 ∧ 2 → (benign) or 3 ∧ 5 → (malignant) or 4 ∧ 2、6 → (benign) or 4 ∧ 1 → (malignant) or 5 ∧ 1、2、4、5、8 → (benign) or 5 ∧ 6、10 → (malignant) or 6 ∧ 3、5 → (benign) 7 ∧ 3、4、6、10 → (malignant) or 8 ∧ 3 → (benign)
12	(1) ∧ (2)	2 ∧ 1 → (benign) or2 ∧ 3 → (malignant)

另外,差转计算及决策树经验推理系统在测试集上的泛化结果见表 7 和表 8。

表 7 差转计算经验推理系统泛化结果

Tab. 7 Generalization results of empirical reasoning systems for set subtraction and rotation calculation

先验类	辨识类		错误率/%
	benign	malignant	
benign	49	2	3.92
malignant	3	24	1.11

表 8 决策树经验推理系统泛化结果

Tab. 8 Generalization results of empirical reasoning systems for decision tree

先验类	辨识类		错误率/%
	benign	malignant	
benign	47	4	3.92
malignant	4	23	1.11

若着重关注结果 malignant,则两种算法的 5 个评估指标结果见表 9。

可以看出差转计算学习能力是较好的,其经验推理系统的泛化能力也与决策树相当,但整体知识的泛化速度要更快。

表 9 算法评估指标结果

Tab. 9 Algorithm evaluation results

算法	指标				
	敏感度/%	查全率/%	查准率/%	错误率/%	*F*1 综合度量
差转计算	96.08	92.31	88.89	6.41	90.57
决策树	92.16	85.19	85.19	5.13	85.19

4 结束语

差转计算算法本质是根据单一条件信息与决策信息在论域中形成的等价类包含关系诱导出概念知识,但存在不能描述不同类型的因素在概念形成过程中的作用的局限。为解决这一问题,本文在因素空间及商集合理论的基础上定义因素析取、合取概念,借此实现数据的粒度变换。实证结论表明本文提出方法是有效的,能够描述不同类型的因素在概念形成过程中的作用,亦能促使差转计算快速收敛,同时挖掘出的知识在泛化过程中决策速率快、决策能力与决策树等同。另外,本文工作仅作为一个研究基础,所涉及的概念转移、因素组合爆炸问题以及知识的修正过程还有待深入研究。

参考文献

[1] 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.

[2] 包研科,茹慧英,金圣军. 因素空间中知识挖掘的一种新算法[J]. 辽宁工程技术大学学报:自然科学版, 2014(33):1144.

[3] 包研科,汪培庄,郭嗣琮. 因素空间的结构与对偶回旋定理[J]. 智能系统学报, 2018, 13(4):168-176.

[4] 包研科,茹惠英. 差转计算的算法与实证[J]. 模糊系统与数学, 2017,13(6):12-45.

[5] 包研科. 泛因素空间与数据科学应用[M]. 北京邮电大学出版社, 2021:115.

[6] 李强. 创建决策树算法的比较研究—ID3,C4.5,C5.算法的比较[J]. 甘肃科学学报,2006(4):88-91.