第 54 卷 第 1 期
2018 年 1 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 54, No. 1
Jan., 2018

DOI:10.13232/j.cnki.jnju.2018.01.016

序贯三支决策的代价敏感分类方法

方 宇[1],闵 帆[1*],刘忠慧[1],杨 新[2]

(1.西南石油大学计算机科学学院,成都,610500;2.四川工商学院,成都,611745)

摘 要:序贯三支决策体现了信息粒化和代价敏感学习的优势,其中信息粒化是人类认知和决策执行的基础,代价则是信息处理涉及的重要因素.提出针对代价敏感学习的序贯三支决策模型.首先,对信息粒化和决策代价之间的关系进行了定义和描述;然后,从序决策过程的视角,利用不同粒度层次的代价矩阵构建了代价函数;最后,为平衡决策结果代价和决策过程代价,提出了两个优化问题,并从理论上阐述了其意义,从实验结果分析上验证了算法的有效性,体现了序贯三支决策在代价敏感分类问题上的优势.

关键词:三支决策,代价敏感学习,粒计算,分 类

中图分类号:TP18 **文献标识码**:A

Sequential three-way decisions based cost-sensitive approach to classification

Fang Yu[1], Min Fan[1*], Liu Zhonghui[1], Yang Xin[2]

(1. School of Computer Science, Southwest Petroleum University, Chengdu, 610500, China;
2. Sichuan Technology and Business University, Chengdu, 611745, China)

Abstract: Sequential three-way decisions take the advantage of information granularity and various types of costs. Information granularity is the basis of human cognition and decision-making, while costs are usually considered as an important information processing related factor. In this paper, we propose a cost-sensitive sequential three-way decision (S3WD) model, the aim of which is to motivate, interpret and implement the three-way decision (3WD) through the notion of information granularity. We are essentially dealing with three difficulties while applying three-way decisions to cost-sensitive learning. The major difficulty is to construct a sequence of multi-level granularities for 3WD. Multi-granulation represents the information granularity with a partial ordering relation, which provides the semantics for the S3WD model. By constructing a family of equivalence classes of particular object with the change of the number of attributes, a series of coarsening-refinement granularities on information system is formalized. Therefore, we can obtain a complete description of a system with multi-granularity. The second difficulty involves the interpretations of evaluation and thresholds in S3WD. We study decision cost related to information granularity with sequential three-way decisions and focus on the formulations of the cost function according to the principle of

基金项目:国家自然科学基金(41604114),西南石油大学科研启航计划(2014QHZ025),西南石油大学第二届中青年骨干教师项目
收稿日期:2017-12-21

* 通讯联系人,E-mail:minfanphd@163.com

coarsening-refinement granulating procedure of information acquirement. The decision cost is composed of the cost of decision result and the cost of decision process, which are two main costs in the cost-sensitive sequential three-way decision model. By utilizing the cost matrix, we construct a reasonable cost structure as evaluation metric for S3WD. The pair of thresholds(α,β)is also obtained. Theoretical analysis indicates that the cost structure has practical and meaningful interpretations of granularity-driven S3WD. The last, crucial difficulty is introducing the objective function to the cost structure. Since there is a trade-off between two costs, we define two optimization problems with the objective of either minimizing the cost of decision result or minimizing the cost of decision process. By solving the optimization problems, the minimal-result-cost S3WD and the minimal-process-cost S3WD are defined as finding optimized S3WD with a minimum cost of decision result or minimum cost of decision process respectively. Two heuristic approaches are designed correspondingly. For minimal-process-cost S3WD, the upper bound of cost of decision result is applied as an input parameter since the optimization problem is a combinational problem. Similar to the minimal-process-cost S3WD, we consider the upper bound of the cost of decision process as a user-defined input parameter for minimal-result-cost S3WD. The two optimization problems represent two risk attitudes of a decision maker in real world scenarios. The experimental results indicate that: 1)with the coarsening-refinement granulating, the cost of the decision result has non-monotonicity subsiding trend, and the cost of the decision process has surging monotonicity trend; 2)with the same number of attributes, the minimal-process-cost S3WD has a higher cost of the decision compared to the minimal-result-cost S3WD; 3)the upper bounds of the costs of the decision result and the upper bounds of the costs of the decision process have directly influence on the decision results; 4)the cost-sensitive S3WD is particularly appropriate for decision-making problems when information is unavailable and costly for on-demand details.

Key words: three-way decisions, cost-sensitive learning, granular computing, classification

代价敏感学习(Cost-sensitive Learning)是数据挖掘和机器学习的重要研究课题,其主要目的是处理在决策过程中产生的各种代价问题.代价敏感学习问题在现实生活生产中具有普适性,例如:医学诊断、机器人、工业生产过程、通信网络故障诊断等等.根据 Hunt *et al*[1]的研究,在代价敏感学习研究中主要有两类代价值得关注:决策误分类代价和对象属性的测试代价.Turney[2]在归纳概念学习研究中对代价进行了分类,对代价敏感学习研究提供了语境.很多研究表明代价敏感学习在决策过程中不仅是重要的而且是必要的.

针对人类的决策认知和规则学习,三支决策[3-5](Three-way Decision,3WD)作为一个重要的决策方法论在近十年得到了蓬勃发展.三支决策主要由两个紧密交织的任务构成:三分和三治.三分指的是把论域划分为三个两两不相交的区域(例如:区域Ⅰ、区域Ⅱ、区域Ⅲ);三治指的是针对在三个不同域中的对象采用不同的治理方式(例如:策略Ⅰ、策略Ⅱ、策略Ⅲ).在不同的研究背景下,很多研究对三分和三治两个任务进行具体的构造和解释,提出了大量三支决策相关的模型和应用.在扩展模型和优化模型上,相关的研究有:决策粗糙集、概率粗糙集、博弈论粗糙集、区间集、模糊区间集、基于三支决策的不完备信息系统、基于统计的三支决策、三支概念格等等[6-8].在应用方面,相关的研究有:临床诊断、论文同行评审、政府和投资决策、文本分类、邮件过滤、推荐系统、聚类分析、人脸识别、属性约简等[9-11].

本文的主要贡献如下:

(1)旨在通过三支决策的驱动、解释和实现,在粒计算[12-13](Granular Computing,GrC)的概念下,提出了代价敏感序贯三支决策模型.

(2)利用不同属性个数反映对系统不同的

描述能力这一特性,构造等价类的集合,实现了对信息系统的粒化,进一步构成了具有序关系的多粒度空间.

(3)研究了序贯三支决策构建多粒度空间过程相关的决策代价.根据信息获取的粒化过程中粒度的粗化-细化原则提出了代价函数;利用代价矩阵,构建了一个合理的代价结构用于评价序贯三支决策,合理地解释了阈值对(α,β).

(4)在粒度由粗到细的变化过程中,决策结果代价具有非单调下降的特点,而决策过程代价具有单调上升的特点,由此提出了两个优化问题,优化目标为:由用户设定一种代价上限,最小化另一种代价,对信息系统的对象进行三分.反映了人们在现实生活中对待风险的两种不同态度.

1 代价敏感序贯三支决策模型

本节主要讨论了粒计算驱使下序贯三支决策的代价敏感方法.首先,从粒计算的概念出发,介绍了序贯三支决策的定义,然后给出了序贯三支决策模型的多粒度空间的构造,最后给出了序贯三支决策中的代价结构,提出代价敏感序贯三支决策模型.

1.1 序贯三支决策 序贯三支决策(Sequential Three-way Decisions,S3WD)理论[14]作为粒计算概念下的产物,其目标是提供一个灵活的机制和方法,使得用户在信息粒化序列过程中作出合适的决策.由于 S3WD 过程需要的时间和信息较少,因此,相比 3WD 而言,S3WD 的优势在于决策过程代价小,决策速度快.例如,通过考察一个对象的描述或者相关信息来判断它的真状态.但是,这样的有效信息往往不完整、不明确;条件集合可能是非标准的;关于对象的描述具有二义性.所以需要延迟决策,只有当信息的细节足够支持决策精度的情况下,才能作出一个明确的决策.这样的决策方式是人类认知和决策的基本原则,这也是 S3WD 模型的基础.

在 S3WD 模型中,假设论域是由独立的元素构成,论域空间有 $n+1$,$n\geqslant 1$ 层粒度,$\{0,1,2,\cdots,n\}$索引集合标识 $n+1$ 层.层序列 n 到 0,标示信息粒从最粗到最细的粒度层.如果对象 $x\in U$ 在每一粒度层下存在一个描述,而对应粒度层的多个描述存在一个全序关系 $\leq$,即:$\mathrm{Des}_0(x)\leq\mathrm{Des}_1(x)\leq\cdots\leq\mathrm{Des}_n(x)$,$\mathrm{Des}_0(x)$是对象 x 最精细的描述,$\mathrm{Des}_n(x)$是最粗糙的描述.针对某一特定层对 U_i,$0\leqslant i-1\leqslant n$ 进行三分,引入评价函数 $v_i(\mathrm{Des}_i(x))$和阈值对(α_i,β_i).对 S3WD 模型给出以下基本定义(包括定义 1 和 2).

定义 1 假设论域 U 有 $n+1$,$n\geqslant 1$ 层粒度,$v_i:U_i\xrightarrow[\mathrm{Des}_i]{}L_i$ 为从 U_i 到全序集$(L_i,\leq_i)$的评价函数,给定一个阈值对$(\alpha_i,\beta_i\in L_i)$且 $\beta_i\prec_i\alpha_i$,$i.e.$,$(\beta_i\prec_i\alpha_i)\wedge\neg(\alpha_i\leq_i\beta_i)$,在特定层 i,$1\leqslant i\leqslant n$ 上,L_i 可以分为三个两两不相交的域:

$$\begin{aligned}&\mathrm{POS}_{(\alpha_i,\beta_i)}(v_i)=\{x\in U_{i+1}\mid v_i(\mathrm{Des}_i(x))\geq_i\alpha_i\}\\&\mathrm{NEG}_{(\alpha_i,\beta_i)}(v_i)=\{x\in U_{i+1}\mid v_i(\mathrm{Des}_i(x))\leq_i\beta_i\}\\&\mathrm{BND}_{(\alpha_i,\beta_i)}(v_i)=\{x\in U_{i+1}\mid\beta_i\prec v_i(\mathrm{Des}_i(x))\prec_i\alpha_i\}\end{aligned}\tag{1}$$

其中,$\mathrm{BND}_{(\alpha_i,\beta_i)}(v_i)$是边界域,边界域中的对象均被延迟决策.随着从低层获取更多的细节信息,边界域会逐渐减小,对象会从 BND 划分到 POS 和 NEG.最终,S3WD 在第 0 层实现简单二支决策.

定义 2 在第 0 层,L_0 可以分为两个不相交的域:

$$\begin{aligned}&\mathrm{POS}_{\gamma_0}(v_0)=\{x\in U_1\mid v_0(\mathrm{Des}_0(x))\geq\gamma_0\}\\&\mathrm{NEG}_{\gamma_0}(v_0)=\{x\in U_1\mid v_0(\mathrm{Des}_0(x))\prec\gamma_0\}\end{aligned}\tag{2}$$

其中,阈值 $\gamma_0\in L_0$ 表示两个域是基于该阈值而划分的.

1.2 序贯三支决策多粒度空间的构建 从粒计算角度定义 S3WD 中的代价函数,首先需要对信息粒化进行合理解释.信息粒代表一个 U 的一个对象子集,它描述了一个系统或问题的子部分,而信息粒的粒度表达了其泛化能力或抽象能力.通过聚集具有相同粒度的信息粒,可以得到一个系统或问题的整体描述,这些粒的集合构成了一个粒度,构建一个粒度的过程称作对系统或问题在特定层的粒化[15].同时,大

的粒是由多个小的粒支撑的.

通常使用等价类$[x]_A$来描述一个对象x,通过增加对象属性个数,其对象的描述能力逐渐增强,这样一个从粗糙到精细的粒化过程反映了多粒度的概念和基本方法.通过不同的等价类,以及在此基础上的等价类划分,能够构建S3WD的多粒度空间.

让$[x]_A$表示为信息粒,$g(A)$为对论域U的划分,其中,A表示为条件属性C的子集.对于决策表,粒化定义如下:

定义 3　给定一个决策表$S=(U,At=C\cup D,\{V_a\mid a\in At\},\{I_a\mid a\in At\})$,假设有$n+1,n\geqslant 1$层粒度,在决策表$S$上的粒化定义为:

$$G=\{g(A_i)\mid A_i\subseteq C\} \tag{3}$$

其中,$g(A_i)$为某一特定相同粒度的信息粒集合对U的划分,$A_i,1\leqslant i\leqslant n$,是条件属性的子集,且满足条件$A_1\subset A_2\subset\cdots\subset A_n\subset A_{n+1}=C$.

例 1　给定一个决策表S,如表1所示,可以根据定义3,对决策系统构建一个多粒度空间.

表 1　决策信息表

Table 1　An exemplary decision table

	a_1	a_2	a_3	a_4	D
x_1	1	0	1	1	d_1
x_2	1	0	1	1	d_1
x_3	0	1	0	1	d_2
x_4	1	1	0	0	d_2
x_5	0	1	1	0	d_2
x_6	1	0	1	0	d_2

表1中共有6个对象、2个决策标签和4个条件属性,从1到4个属性递增方式,在不同属性集合下,构造等价类.让$g(A_i),0\leqslant i-1\leqslant 3$为第$i$层粒度下,信息粒的集合划分.由此可以构造4层粒度空间,例如,假设一种简单的属性个数增加过程为$A_0=\{a_1\},A_1=\{a_1,a_2\},A_2=\{a_1,a_2,a_3\},A_3=\{a_1,a_2,a_3,a_4\}$,最粗糙的粒度则为$g(A_0)$,最精细的粒度为$g(A_3)$.粒化情况如下:

第3层,$g(A_0)=g(\{a_1\})=\{\{x_1,x_2,x_4,x_6\},\{x_3,x_5\}\}$;

第2层,$g(A_1)=g(\{a_1,a_2\})=\{\{x_1,x_2,x_6\},\{x_3,x_5\},\{x_4\}\}$;

第1层,$g(A_2)=g(\{a_1,a_2,a_3\})=\{\{x_1,x_2,x_6\},\{x_3\},\{x_5\},\{x_4\}\}$;

第0层,$g(A_3)=g(\{a_1,a_2,a_3,a_4\})=\{\{x_1,x_2\},\{x_3\},\{x_5\},\{x_4\},\{x_6\}\}$.

例1中,可以观察到两两相邻的层存在一个偏序关系,且这个偏序关系存在传递性.

定义 4　给定一个决策表S,有$n+1,n\geqslant 1$层粒度,P_i表示第i层粒化,Q_j表示第j层粒化,$0\leqslant i-1\leqslant j-1\leqslant n$,如果满足以下关系$Q\preceq P$则成立:

$$\forall Q\in Q_j,\exists P\in P_j\quad Q\subseteq P \tag{4}$$

其中,P,Q分别是P_i,Q_j粒化下的粒集合.

例1中,$g(A_0),g(A_1),g(A_2),g(A_3)$是构成4层粒度的粒化过程.根据定义4,可以断定$g(A_1)\preceq g(A_0)$:$g(A_0)$有两个粒$\{x_1,x_2,x_4,x_6\}$和$\{x_3,x_5\}$,$g(A_1)$有三个粒$\{x_1,x_2,x_6\}$,$\{x_3,x_5\}$和$\{x_4\}$.对于在$g(A_0)$中的任意粒$Q$,能够找到一个$g(A_1)$中的粒$P$,使得$Q\subseteq P$.例如,$\{x_1,x_2,x_6\}\in\{x_1,x_2,x_4,x_6\}$或$\{x_4\}\in\{x_1,x_2,x_4,x_6\}$.同样地,可以得到$g(A_2)\preceq g(A_1)$,$g(A_3)\preceq g(A_2)$.因此,很容易证明$g(A_3)\preceq g(A_2)\preceq g(A_1)\preceq g(A_0)$这样的关系(证明略),故定义4中$Q\preceq P$具有传递性.

对于决策表的多粒度空间构建和解释,定义3和定义4给出了基本语义.而对于S3WD多粒度层的构建和理解正是在此基础上引入代价结构和阈值对而成立的.

1.3　序贯三支决策模型代价结构设计　在S3WD模型中主要有两种代价,一种是决策结果代价(可看作对象误分类而产生的风险),如表2所示;另一种代价为决策过程代价(可看作描述对象或获得对象属性值而产生的代价).但在大多数3WD研究中,决策过程代价往往被忽略了.代价结构设计中,决策过程代价是一个关键因素,只有对其给定良好的定义,S3WD在决策方法中的优势才能得以体现.

表 2 决策结果代价矩阵

Table 2 Decision result cost matrix

	$X(P)$	$X^C(N)$
a_P	$\lambda_{PP}=\$0$	$\lambda_{PN}=\$5$
a_B	$\lambda_{BP}=\$1$	$\lambda_{BN}=\$2$
a_N	$\lambda_{NP}=\$4$	$\lambda_{NN}=\$0$

在提出的 S3WD 模型中,在做出明确的决策前,(决策过程中)有一系列的属性测试和延迟决策,对应的代价为测试代价和延迟代价,如表 3 所示,关于代价函数在不同粒度层之间的变化过程可以标识为两两相邻层之间的可重复的序列操作.

表 3 决策过程代价向量

Table 3 Decision process cost vector

	a_1	a_2	a_3	a_4
$tc(a)$	\$3	\$5	\$4	\$5
$dc(a)$	2 mins	8 mins	10 mins	6 mins

定义 5 给定一个决策表 S,有 $n+1,n\geqslant 1$ 层粒度,在 S 上的 S3WD 代价结构可定义为:

$$COST=(COST_R,COST_P) \tag{5}$$

其中,$COST_R$ 和 $COST_P$ 分别表示决策结果代价和决策过程代价.

假设在 S 上进行粒化,有一个属性集合序为:$A_0,A_1,\cdots,A_n$;其中,$g(A_i),0\leqslant i-1\leqslant n$ 表示第 $n-i$ 层上的信息粒集合,那么,在该层上,对象 x 的决策结果代价为:

$$COST_R=\lambda(a(x)|g(A_i)) \tag{6}$$

信息粒集合 $g(A_i)$ 的决策结果代价为:

$$COST_R=\sum_{x\in U_i}\lambda(a(x)|g(A_i)) \tag{7}$$

其中,$g(A_i)$ 与定义 3 中相同,$a(x)$ 表示为对象 x 做出的决策,即 a_P,a_N,a_B 中的某一个,$\lambda\in\{\lambda_{PP},\lambda_{PN},\lambda_{BP},\lambda_{BN},\lambda_{NP},\lambda_{NN}\}$.

对象 x 的决策过程代价为:

$$COST_P(x)=(tc(x),dc(x)) \tag{8}$$

其中,$tc(x)$ 和 $dc(x)$ 分别表示属性测试代价和对象延迟代价,其代价函数表示为:

$$tc(x)=tc(A_i)=\sum_{a\in A_i}tc(a) \tag{9}$$

$$dc(x)=dc(A_i)=\max_{a\in A_i}dc(a) \tag{10}$$

值得注意的是,在一个粒度下,多个信息粒的属性是同时测试的,考虑对象的属性之间是相互独立的,因此可以得到在第 i 层上,信息粒结集合 $g(A_i)$ 的决策过程代价为(假定 S3WD 在第 n 层,即最粗层进行划分时,令 $U_n=\mathrm{BND}_{(\alpha_n,\beta_n)}$):

$$COST_P=\Big(\sum_{x\in BND_{(\alpha_{i+1},\beta_{i+1})}}tc(x),\sum_{x\in BND_{(\alpha_{i+1},\beta_{i+1})}}dc(x)\Big) \tag{11}$$

2 代价敏感序贯三支决策模型算例

本节主要通过一个算例演示 S3WD 的决策过程.代价结构的构建为 S3WD 模型提供了语义和解释.使得代价敏感 S3WD 模型具有理论意义和应用场景.

在例 1 中,$g(A_0),g(A_1),g(A_2),g(A_3)$ 构成了 4 层粒度.假设代价函数由表 2 和表 3 给定,为了简化计算复杂度,设每一层粒度下的决策结果代价矩阵相同.由此,可计算得到阈值 $\alpha=0.75,\beta=0.4$.阈值对具体的计算过程在有关决策粗糙集、三支决策的文献中均有详细描述[7-8,11,14],本文不再赘述.现在,根据式(7),计算决策表(如表 1 所示)每一粒度层下的决策结果代价.

(1)在第 3 层,对于对象 x_1,其条件概率为:

$$Pr(D=d_1|[x_1]_{A_0}=\frac{|\{x_1,x_2\}\bigcap\{x_1,x_2,x_4,x_6\}|}{|\{x_1,x_2,x_4,x_6\}|}=0.5$$

$$\alpha>Pr(D=d_1|[x_1]_{A_0})>\beta$$

因此,根据定义 1,$x_1\in\mathrm{BND}_{(\alpha_3,\beta_3)}(D)$;同样地,可以对这一层所有对象进行判断,由此,三个域可划分为:

$$\mathrm{POS}_{(\alpha_3,\beta_3)}(D)=\varnothing$$

$$\mathrm{NEG}_{(\alpha_3,\beta_3)}(D)=\{x_3,x_5\}$$

$$\mathrm{BND}_{(\alpha_3,\beta_3)}(D)=\{x_1,x_2,x_4,x_6\}$$

根据式(7):

$$COST_R(g(A_0))=0\times(\$5\times0.5)+4\times(\$1\times0.5+\$2\times0.5)+2\times(\$4\times0.5)=\$10$$

(2)在第 2 层,已划分到域 POS 和 NEG 的对象不再考察,只需要对第 3 层 BND 中的对象进一步判断,执行在第 3 层中同样的计算,三个域可划分为:

$$\mathrm{POS}_{(a_2,\beta_2)}(D)=\varnothing$$
$$\mathrm{NEG}_{(a_2,\beta_2)}(D)=\{x_3,x_4,x_5\}$$
$$\mathrm{BND}_{(a_2,\beta_2)}(D)=\{x_1,x_2,x_6\}$$
$$COST_R(g(A_1))=\$6.67$$

(3)在第 1 层,三个域可划分为:

$$\mathrm{POS}_{(a_1,\beta_1)}(D)=\varnothing$$
$$\mathrm{NEG}_{(a_1,\beta_1)}(D)=\{x_3,x_4,x_5\}$$
$$\mathrm{BND}_{(a_1,\beta_1)}(D)=\{x_1,x_2,x_6\}$$
$$COST_R(g(A_2))=\$4$$

(4)在第 0 层,三个域可划分为:

$$\mathrm{POS}_{(a_0,\beta_0)}(D)=\{x_1,x_2\}$$
$$\mathrm{NEG}_{(a_0,\beta_0)}(D)=\{x_3,x_4,x_5,x_6\}$$
$$\mathrm{BND}_{(a_0,\beta_0)}(D)=\varnothing$$
$$COST_R(g(A_3))=\$4$$

根据式(11),可以计算每一粒度层下的决策过程代价.

(1)在第 3 层,有 1 个属性$\{a_1\}$进行了测试,6 个对象进行了判断,因此,$tc(g(A_0))=6\times\sum_{a\in A_0}tc(a)=6\times tc(\{a_1\})=6\times\$3=\$18$,$dc(g(A_0))=6\times\max_{a\in A_0}dc(a)=6\times dc(\{a_1\})=6\times 2\ \mathrm{min}=12\ \mathrm{min}$;$COST_P(g(A_0))=(\$18,12\ \mathrm{min})$.

(2)在第 2 层,有 2 个属性$\{a_1,a_2\}$进行了测试,4 个对象进行了判断,因此,$tc(g(A_1))=4\times\sum_{a\in A_1}tc(a)=4\times tc(\{a_1,a_2\})=4\times(\$3+\$5)=\$32$,$dc(g(A_1))=4\times\max_{a\in A_1}dc(a)=4\times dc(\{a_1,a_2\})=4\times 8\ \mathrm{min}=32\ \mathrm{min}$;$COST_P(g(A_1))=(\$32,32\ \mathrm{min})$.

(3)在第 1 层,有 3 个属性$\{a_1,a_2,a_3\}$进行了测试,3 个对象进行了判定,因此,$tc(g(A_2))=3\times\sum_{a\in A_2}tc(a)=3\times tc(\{a_1,a_2,a_3\})=3\times(\$3+\$5+\$4)=\$36$,$dc(g(A_2))=3\times\max_{a\in A_2}dc(a)=3\times dc(\{a_1,a_2,a_3\})=3\times 10\ \mathrm{min}=30\ \mathrm{min}$;$COST_P(g(A_2))=(\$36,30\ \mathrm{min})$.

(4)在第 0 层,有 4 个属性$\{a_1,a_2,a_3,a_4\}$进行了测试,3 个对象进行了判定,因此,$tc(g(A_3))=3\times\sum_{a\in A_3}tc(a)=3\times tc(\{a_1,a_2,a_3,a_4\})=3\times(\$3+\$5+\$4+\$5)=\$51$,$dc(g(A_3))=3\times\max_{a\in A_3}dc(a)=3\times dc(\{a_1,a_2,a_3,a_4\})=3\times 10\ \mathrm{min}=30\ \mathrm{min}$;$COST_P(g(A_3))=(\$51,30\ \mathrm{min})$.

最终获得在 4 个不同粒度下三个域的划分情况,以及对应的决策代价,如表 4 所示.

表 4　序贯三支决策算例结果

Table 4　Computing results of S3WD

Level Index	POS	BND	NEG	$COST_R$	$COST_P$
3	$\varnothing$	$\{x_1,x_2,x_4,x_6\}$	$\{x_3,x_5\}$	\$10	(\$18,12 mins)
2	$\varnothing$	$\{x_1,x_2,x_6\}$	$\{x_3,x_4,x_5\}$	\$6.67	(\$32,32 mins)
1	$\varnothing$	$\{x_1,x_2,x_6\}$	$\{x_3,x_4,x_5\}$	\$4	(\$36,30 mins)
0	$\{x_1,x_2\}$	$\varnothing$	$\{x_3,x_4,x_5,x_6\}$	\$4	(\$51,30 mins)

3　决策代价

本节在探讨决策结果代价和决策过程代价的关系、各自特点的基础上,提出了两个优化问题,并对每个优化问题给出了解释性算法.

3.1　决策结果代价与决策过程代价　从第 2 节的算例中可以观察到随着粒度的变细,$COST_R$ 总体呈现下降的趋势,但如果进一步分析发现有些属性是可约简的,因此,即使考察了这些属性,仍然不能提高分类精度,至此导致 $COST_R$ 具有非单调性,非单调性已经在部分文献[14,16]中进行了充分证明,本文不再赘述.

而对于 $COST_P$ 总体呈现单调增加的趋势,即:$COST_P(g(A_n))>COST_P(g(A_{n-1}))>\cdots>COST_P(g(A_0))$. 有两个原因可以解释这一特性:(1)测试代价随着属性个数增加而增加.(2)延迟代价随着层的增加而累加.

总的来说,决策过程代价和决策过程代价是此消彼长的关系. 决策者可以在粗粒度下做出快速决策且决策过程代价小,但决策结果代价大;相反的,决策者可以在细粒度下做出决策,决策结果代价小且分类精度高,但决策过程代价大. 因此,在这两种代价之间寻找一个平衡点成为了决策者有效利用 S3WD 方法的关键.

3.2 最小决策结果代价序贯三支决策模型

在一些现实应用中,相比决策过程代价,一般更关注决策结果代价,从而避免高风险决策. 例如,在疑似癌症患者的诊断过程中,患者做了 X 光检查,但检查获得的信息不足以让医生做出判断,为了避免误诊(风险大)患者需要进一步做 MRI 检查,由此导致的决策过程代价增加.

在这种情况下,优化的目标就是最小决策结果代价. 一个可行的方法是决策者设定决策过程代价上限的基础上,在 S3WD 过程中找到最小决策结果代价的粒度层下的对象划分(即,在该层上的 3WD).

定义 6 给定一个决策表 S,有 $n+1$,$n\geqslant 1$ 层粒度,设定决策过程代价上限 $COST_P^{UB}$,$P\subseteq G$ 是 S3WD 中,最小决策结果代价相关粒度层下的对象划分,当且仅当满足以下情况成立:

(1) $P=\mathrm{argmin}_{P\subseteq G}\{COST_R\}$;

(2) $\forall Q\leq P$, $\sum_j^n COST_P>COST_P^{UB}$.

其中,G 的含义同定义 3,Q,$P\subseteq G$ 分别表示在 j,i 层上的三支决策,$0\leqslant i<j\leqslant n$.

算法 1 最小决策结果代价 S3WD 算法

输入:决策表 S,$n+1$,$n\geqslant 1$ 层粒度,代价结构 $COST=(COST_R,COST_P)$,决策过程代价上限 $COST_P^{UB}$;

输出:min$COST_R$ 和相关粒度层下的三个域 POS,BND,NEG.

①begin

POS=∅;BND=∅;NEG=∅;

②$i=n+1$;$U_{n+1}=U$;

③Compute the pair of thresholds(α,β);

④while $\sum_j^n COST_P\leq COST_P^{UB}$ and $i>0$ do

⑤$i=i-1$;

⑥foreach $[x]_{A_i}\subseteq U_i$ do

⑦if $Pr(D|[x]_{A_i})\geqslant\alpha$ then POS=POS∪$[x]_{A_i}$;

⑧if $Pr(D|[x]_{A_i})\leqslant\beta$ then NEG=NEG∪$[x]_{A_i}$;

⑨BND=U_i−POS−NEG;

⑩U_i=BND;

⑪Compute $COST_P$ and $COST_R$;

⑫if $U_i\neq\varnothing$ and $\sum_j^n COST_P\leq COST_P^{UB}$ then

⑬foreach$[x]_{A_0}\subseteq U_i$ do

⑭if $Pr(D|[x]_{A_0})\geqslant\gamma_0$ then POS=POS∪$[x]_{A_0}$;

⑮if $Pr(D|[x]_{A_i})<\gamma_0$ then NEG=NEG∪$[x]_{A_0}$;

⑯Compute $COST_R$

⑰return $COST_R$,POS,BND,NEG.

3.3 最小决策过程代价序贯三支决策模型

相对于最小决策结果代价 S3WD 模型,决策过程代价 S3WD 模型也有相关的应用场景,由于文章篇幅原因,以下仅给出定义和算法.

定义 7 给定一个决策表 S,有 $n+1$,$n\geqslant 1$ 层粒度,设定决策结果代价上限 $COST_R^{UB}$,$P\subseteq G$ 是 S3WD 中,最小决策过程代价相关粒度层下的对象划分,当且仅当满足以下情况成立:

(1) $P=\mathrm{argmin}_{P\subseteq G}\{COST_P\}$;

(2) $\forall Q\leq P$,$COST_R(Q)\geq COST_R^{UB}$.

其中,G 的含义同定义 3,Q,$P\subseteq G$ 分别表示在 j,i 层上的三支决策,$0\leqslant i<j\leqslant n$.

算法 2 最小决策过程代价 S3WD 算法

输入:决策表 S,$n+1$,$n\geqslant 1$ 层粒度,代价结构 $COST=(COST_R,COST_P)$,决策结果代价上限 $COST_R^{UB}$;

输出:min$COST_P$ 和相关粒度层下的三个域 POS,BND,NEG.

①begin

POS=∅;BND=∅;NEG=∅;

②$i=n+1$;$U_{n+1}=U$;

③Compute the pair of thresholds(α,β);

④while $COST_{R_{i+1}}\geqslant COST_R^{UB}$ and $i>0$ do

⑤foreach$[x]_{A_i}\subseteq U_i$ do

⑥if $Pr(D|[x]_{A_i})\geqslant\alpha$ then POS=POS$\cup[x]_{A_i}$;
⑦if $Pr(D|[x]_{A_i})\leqslant\beta$ then NEG=NEG$\cup[x]_{A_i}$;
⑧BND=U_i−POS−NEG;
⑨U_i=BND;
⑩$i=i-1$;
⑪Compute $COST_P$ and $COST_R$;
⑫if $U_i\neq\varnothing$ and $COST_{R_0}\geqslant COST_R^{UB}$ then
⑬foreach$[x]_{A_0}\subseteq U_i$ do
⑭if $Pr(D|[x]_{A_0})\geqslant\gamma_0$ then POS=POS$\cup[x]_{A_0}$;
⑮if $Pr(D|[x]_{A_i})<\gamma_0$ then NEG=NEG$\cup[x]_{A_0}$;
⑯Compute $COST_P$
⑰return $COST_P$,POS,BND,NEG.

3.2 和 3.3 节提出的两个模型反映了决策者决策过程中的两种风险态度和两种决策策略,运用这两个模型和相应的算法,可以在现实生活中平衡两种代价,从而做出符合实际的决策.

4　实　验

选用 4 个 UCI 数据集运用于算法的验证,假设每个数据集的决策属性为两类且没有缺失值.数据集基本信息以及决策结果代价和决策过程代价设定,如表 5 所示.

算法 1 和 2 在每个数据集上运行了 3000 次,用卡方值 χ^2 验证算法的有效性,使用了近似质量 $\gamma=\dfrac{|X-\mathrm{POS}|+|X^C-\mathrm{NEG}|}{|U|}$ 评价了分类效果,使用粒度的层数来验证序贯三支决策在分类问题上的有效性.使用了 Matlab R2016a 作为算法实现的平台.实验统计结果如表 6 所示.

表 5　UCI 数据集基本信息和代价上限设定

Table 5　Briefings of data sets and settings for cost upper bound

Data set	$\|U\|$	$\|C\|$	$\|X\|$	$COST_R^{UB}$	$COST_P^{UB}$
Acute	120	6	30	60	[80,90]
Breast	699	9	241	200	[200,300]
Voting	435	16	168	100	[150,200]
Pima	768	8	268	200	[150,150]

表 6　卡方、近似质量和粒度层数的平均值以及标准差

Table 6　Means and standard deviations of χ^2, approximate quality and levels of granularity

Data set	算法 1			算法 2		
	γ	χ^2 statistic	Level of granularity	γ	χ^2 statistic	Level of granularity
Acute	0.30±0.22	58.28±40.96	2.28±0.71	0.37±0.16	47.17±32.01	2.15±0.84
Breast	0.09±0.10	502.93±85.77	4.91±1.40	0.10±0.10	497.47±79.51	4.12±1.65
Voting	0.07±0.06	250.96±46.62	4.89±1.48	0.08±0.06	255.55±43.22	8.74±4.36
Pima	0.17±0.07	545.67±46.62	4.46±0.82	0.18±0.07	546.06±80.84	4.45±1.28

5　结　论

本文从粒计算的角度考察了代价敏感学习和 S3WD 过程,提出了 S3WD 下的代价敏感分类模型和分类方法.讨论了 S3WD 模型中的两类代价以及之间的关系,在此基础上提出了两个优化问题并提出了相关的算法.理论和实验结果分析表明:代价敏感 S3WD 在实际应用中更有优势.但本文在多层粒度下阈值对该如何变化并没有深入讨论,这将成为后续研究工作的重点.

参考文献

[1] Hunt E B, Marin J, Stone P J. Experiments in induction. Oxford: Academic Press, 1966, 98－120.

[2] Turney P D. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of Artificial Intelligence Research, 1995, 2(1): 369－409.

[3] Yao Y Y. An outline of a theory of three-way decisions // Proceedings of the 8th International Conference Rough Sets and Current Trends in Computing 2012. Springer Berlin Heidelberg, 2012: 1－17.

[4] Yao Y Y. Three-way decisions and cognitive computing. Cognitive Computation, 2016, 8(4): 543－554.

[5] Yao Y Y. The superiority of three-way decisions in probabilistic rough set models. Information Sciences, 2011, 181(6): 1080－1096.

[6] Deng X F, Yao Y Y. Decision-theoretic three-way approximations of fuzzy sets. Information Sciences, 2014, 279: 702－715.

[7] Liu D, Li T R, Ruan D. Probabilistic model criteria with decision-theoretic rough sets. Information Sciences, 2011, 181(17): 3709－3722.

[8] Liu D, Liang D C, Wang C C. A novel three-way decision model based on incomplete information system. Knowledge-Based Systems, 2016, 91: 32－45.

[9] Yao J T, Azam N. Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets. IEEE Transactions on Fuzzy Systems, 2015, 23(1): 3－15.

[10] Yu H, Zhang C, Wang G Y. A tree-based incremental overlapping clustering method using the three-way decision theory. Knowledge-Based Systems, 2016, 91: 189－203.

[11] Jia X Y, Liao W H, Tang Z M, *et al*. Minimum cost attribute reduction in decision-theoretic rough set models. Information Sciences, 2013, 219: 151－167.

[12] Fujita H, Li T R, Yao Y Y. Advances in three-way decisions and granular computing. Knowledge-Based Systems, 2016, 91: 1－3.

[13] Yao Y Y, Deng X F. A granular computing paradigm for concept learning // Ramanna S, Jain L, Howlett R. Emerging Paradigms in Machine learning. Springer Berlin Heidelberg, 2013: 307－326.

[14] Yao Y Y. Granular computing and sequential three-way decisions // Lingras P, Wolski M, CornelisC, *et al*. Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2013: 16－27.

[15] Yao Y Y. A triarchic theory of granular computing. Granular Computing, 2016, 1(2): 145－157.

[16] Min Fan, Zhu W. A competition strategy to cost-sensitive decision trees // Li T R, Nguyen H S, Wang G Y, *et al*. Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2012: 359－368.

(责任编辑　杨可盛)