DOI: 10.11992/tis.201706047
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180419.1332.008.html

概率粗糙集三支决策在线快速计算算法研究

徐健锋[1,2], 何宇凡[1], 汤涛[1], 赵志宾[1,2]
（1. 南昌大学 软件工程系, 江西 南昌 330029; 2. 同济大学 计算机科学与技术系, 上海 201804）

摘　要: 随着大数据和物联网技术的不断发展, 动态在线计算已经成为了一种常见的计算模式, 在动态在线计算中进行不确定问题的推理和求解是一项具有挑战性的新议题。概率粗糙集三支决策理论是一种处理不确定性知识挖掘的有效工具, 根据在线计算模式中数据同步增减的动态特点, 提出了一种概率粗糙集三支决策的在线计算方法。首先, 以内存滑动窗口模式对在线动态计算的数据变化特点进行理论建模; 然后, 根据上述模型中在线计算的数据变化模式, 推导出不同类型数据变化模式下的三支决策条件概率及三支区域的变化规律; 最后, 提出了一种新型在线快速计算算法, 其获取的三支决策规则与经典概率三支决策算法是等效的。通过与经典三支决策计算算法的多组对比实验, 验证了提出的在线快速计算算法的高效性与稳定性。

关键词: 三支决策; 粗糙集; 条件概率; 在线计算; 不确定; 动态计算; 粒计算
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)05−0741−10

中文引用格式: 徐健锋, 何宇凡, 汤涛, 等. 概率粗糙集三支决策在线快速计算算法研究[J]. 智能系统学报, 2018, 13(5): 741–750.
英文引用格式: XU Jianfeng, HE Yufan, TANG Tao, et al. Research on a fast online computing algorithm based on three-way decisions with probabilistic rough sets[J]. CAAI transactions on intelligent systems, 2018, 13(5): 741–750.

Research on a fast online computing algorithm based on three-way decisions with probabilistic rough sets

XU Jianfeng[1,2], HE Yufan[1], TANG Tao[1], ZHAO Zhibin[1,2]
(1. Department of Software Engineering, Nanchang University, Nanchang 330029, China; 2. Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

Abstract: With the continuous development of big data and IoT (Internet of Things), dynamic online computation has become a common computing pattern; however the field of dynamic online computation faces challenges in deducing and solving uncertainty problems. A three-way decision theory with probabilistic rough set method is an efficient tool for mining uncertain knowledge; thus a dynamic online computing approach of three-way decision theory with probabilistic rough set is proposed in this paper, in accordance with the features of data dynamic synchronization. First, a data model is established to describe the inherent features of dynamic online computation via memory sliding window mode. In terms of the variational features of dynamic online computation of the above model, a three-way decision conditional probability and the change rule of three-way area are deduced as diverse variational patterns of data. Finally, a novel algorithm of online rapid computation is proposed. The obtained three-way decision rule is identical with the three-way decision algorithm of classic probability. By comparison with the classic three-way decision algorithm through multiple experiments, the proposed online rapid computation algorithm is confirmed to have high efficiency and stability.

Keywords: three-way decisions; rough sets; conditional probability; online computing; uncertain; dynamic calculation; granular computing

随着大数据与互联网加时代的到来, 动态流数据成为了一种主流的数据类型, 当前已经被广泛地应用于金融股票交易、天气和环境监测、计算机网络监控等众多领域[1]。在上述应用中, 在

收稿日期: 2017−06−12.　**网络出版日期:** 2018−04−19.
基金项目: 国家自然科学基金项目 (61763031, 61673301); 上海市自然科学基金项目 (14ZR1442600); 江西省研究生创新专项资金项目 (YC2016-S053).
通信作者: 徐健锋. E-mail: jianfeng_x@ncu.edu.cn.

线计算[2-3]是动态流数据的主要计算形式，其主要特点是实时数据快速地加载入内存，而 CPU 需要对内存中的实时数据实施快速计算，并且实时反馈计算结果。而传统的离线计算[4]方式需要外部存储器积淀数据后再进行科学计算，对于数据实时性和流动性的要求相对较低，因此传统计算方法不能完全适应在线计算的新要求。研究更加高效的在线计算方法已经成为了当前大数据领域的一项重要课题。

近年来，以 SPARK[5]为代表的内存计算平台的推出与发展，较快推动了动态在线计算的发展。不确定性问题是机器学习领域的经典难题，如何在在线计算过程中进行不确定性问题的动态推理及求解也逐渐成为一项具有挑战性的新议题。

三支决策理论[6]是基于粒计算粗糙集理论提出和发展起来的一种处理不确定、不完整信息决策的新方法。其初始目的是为粗糙集理论中的 3 个分类区域，即正域、负域和边界域，提供合理的决策语义解释。近年来随着该理论的深入研究与发展，三支决策被认为是在信息不足或者获取足够信息的代价较高背景下的合理选择。该理论与人类的认知习惯相似，具有认知方面的优势，近年来，三支决策理论在垃圾邮件过滤[7]、文本挖掘[8]、图像识别[9]、属性约简[10]、聚类[11]等应用领域[12-14]也都取得了广泛的成果。

目前主要代表性的理论成果包括：基于代价敏感度量的决策粗糙集三支决策[15]、模糊集三支决策[16]、区间集三支决策[17]、概率粗糙集三支决策[18]、博弈粗糙集三支决策[19]、信息熵三支决策[20]、GINI 系数三支决策[21]等。

同时针对动态数据环境下的动态计算也是另一项主要研究内容。文献[22]首次提出一种离线动态计算方法，其对更新后的数据对象进行重新计算，算法执行效率不高。为了提高动态计算的效率，许多学者研究了粗糙集及三支决策的离线增量学习方法。其主要成果有：Liu 等[23]提出了一种基于矩阵的动态不完备信息系统的增量方法；Li 等[24]通过分析动态数据库中新数据的不断更新，提出了基于粗糙集理论的增量学习方法；Zhang 等[25]等给出了邻域粗糙集模型下多对象增加删除时近似集快速更新的方法；Luo 等[26]考虑到信息系统中数据对象动态插入，讨论介绍了概率粗糙集模型中条件概率的动态估计策略，进而给出了概率三支近似的增量更新算法。然而研究表明，上述动态学习研究都是以离线计算为研究背景，而以在线计算为背景的动态三支决策研究较少。

本文以上述三支决策动态计算研究为基础，系统性地研究了在线计算中动态数据变化形式及动态决策的变化规律，提出一种有别于传统经典计算的三支决策在线快速计算学习方法，并进行实验验证。

1　相关工作

1.1　概率粗糙集三支决策模型

概率粗糙集是构造三支决策的基础原型，首先我们回顾一下概率粗糙集三支决策的相关基本理论[20]。

信息系统IS是一个四元组，$\mathrm{IS}=(U,A,V,f)$，其中U代表论域中对象x的集合；$A=R\cup D$代表属性集合，R 为条件属性集合 ($U/R=\{R_1,R_2\cdots,R_m\}$为 R 属性确定的不可区分关系形成的等价类集合)，D 为决策属性集合 ($U/D=\{D_1,D_2\cdots,D_n\}$为 D 属性确定的不可区分关系形成的等价类集合)；V 代表 A 中各属性的取值范围；f代表从对象到属性取值的信息函数，即$f:U\times A\to V$。

定义 1　在信息系统IS中，设D_j为论域 U 的子集，即$D_j\subseteq U$，给定一个等价类$[x]\subseteq U$属于D_j的条件概率可定义为$\Pr(D_j|[x])=\dfrac{|D_j\cap[x]|}{|[x]|}$。

注：$|\cdot|$表示集合的基数。

定义 2　设D_j为论域 U 的子集，即$D_j\subseteq U$，给定一对阈值 (α,β)，并且满足$0\leqslant\beta<\alpha\leqslant 1$，则 (α,β)-正域、负域和边界域可以定义为

$$\mathrm{POS}_{(\alpha,\bullet)}(D_j)=\{x\in U|\Pr(D_j|[x])\geqslant\alpha\}$$
$$\mathrm{NEG}_{(\bullet,\beta)}(D_j)=\{x\in U|\Pr(D_j|[x])\leqslant\beta\}$$
$$\mathrm{BND}_{(\alpha,\beta)}(D_j)=\{x\in U|\beta<\Pr(D_j|[x])<\alpha\}$$

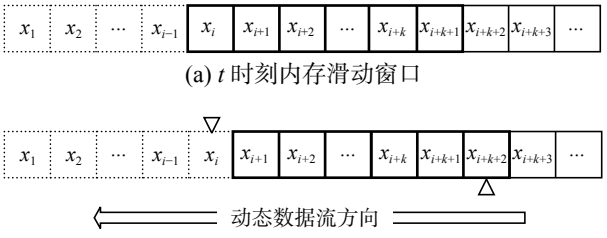
通过定义 2 中的 (α,β)-正域、负域和边界域，可分别生成相应的接受、拒绝和延迟决策规则。

1.2　三支决策在线快速计算模型

滑动窗口技术[27-28]可以有效处理动态流数据下的数据挖掘与知识更新问题。在线计算的动态特点是，随着新数据不断进入，同时由于内存空间的限制，旧数据也被不断被删除，内存中包含的计算数据既增又减。由此，三支决策在线计算的数据变化模式可以用滑动窗口的方式进行描述。

给定 t 时刻的决策信息系统表：$\mathrm{IS}^{(t)}=\{U^{(t)},C^{(t)}\cup D^{(t)}\}$，其中$U^{(t)}/C^{(t)}=\left\{R_1^{(t)},R_2^{(t)},R_3^{(t)},\cdots,R_m^{(t)}\right\}$，$U^{(t)}/D^{(t)}=\left\{D_1^{(t)},D_2^{(t)},D_3^{(t)},\cdots,D_n^{(t)}\right\}$。将决策表空间看作内存空间，即当前决策表中为内存中实时计算的数据。当 t+1 时刻有新的数据对象动态进出决策表时，内存空间发生数据动态变化，其变化机制如下。

设 t 时刻内存滑动窗口如图 1(a) 所示, 其中粗实线框 (即 $x_i, x_{i+1}, \cdots, x_{i+k+1}$ 数据对象) 为当前 t 时刻内存滑动窗口空间, 长度为 k; 虚线框 (即 $x_1, x_2, \cdots, x_{i-1}$ 数据对象) 为历史遗弃数据; 细实线框 (即 $x_{i+k+2}, x_{i+k+3}, \cdots$ 数据对象) 为待移入内存数据。如图 1(b) 所示, t+1 时刻单个对象 x_{i+k+2}(图中上三角标识) 移入内存滑动窗口中, 同时前一时刻最陈旧的决策对象 x_{i-1}(图中下三角标识) 被移出, 内存滑动窗口发生更新。



(a) t 时刻内存滑动窗口

(b) t+1 时刻内存滑动窗口

图 1　在线计算内存数据变化机制模型

Fig. 1　Sliding windows online computing model on memory data

基于三支决策理论以及上述在线内存数据变化模型, 可以获取如下动态单对象三支决策在线内存计算变化模型。

给定 t+1 时刻决策表 $\mathrm{IS}^{(t+1)}$, 当有新数据对象集 $\{\bar{x}\}$ 移入, 同时必然有旧数据对象集 $\{\underline{x}\}$ 移出, 则更新后的决策信息系统中条件等价类和决策类可能分别发生如下变化:

$$R_i^{(t+1)} = \begin{cases} R_i^{(t)} - \{\underline{x}\}, (x,\underline{x}) \in R_i^{(t+1)}, (x,\bar{x}) \notin R_i^{(t+1)}, 1 \leqslant i \leqslant m \\ R_i^{(t)} \cup \{\bar{x}\}, (x,\underline{x}) \notin R_i^{(t+1)}, (x,\bar{x}) \in R_i^{(t+1)}, 1 \leqslant i \leqslant m \\ R_i^{(t)} \cup \{\bar{x}\} - \{\underline{x}\}, (x,\underline{x}) \in R_i^{(t+1)}, (x,\bar{x}) \in R_i^{(t+1)}, 1 \leqslant i \leqslant m \\ R_i^{(t)}, (x,\underline{x}) \notin R_i^{(t+1)}, (x,\bar{x}) \notin R_i^{(t+1)}, 1 \leqslant i \leqslant m \end{cases} \tag{1}$$

$$D_j^{(t+1)} = \begin{cases} D_j^{(t)} - \{\underline{x}\}, (x,\bar{x}) \in D_j^{(t+1)}, (x,\underline{x}) \notin D_j^{(t+1)}, 1 \leqslant j \leqslant n \\ D_j^{(t)} \cup \{\bar{x}\}, (x,\bar{x}) \in D_j^{(t+1)}, (x,\underline{x}) \notin D_j^{(t+1)}, 1 \leqslant j \leqslant n \\ D_j^{(t)} \cup \{\bar{x}\} - \{\underline{x}\}, (x,\bar{x}) \in D_j^{(t+1)}, (x,\underline{x}) \in D_j^{(t+1)}, 1 \leqslant j \leqslant n \\ D_j^{(t)}, (x,\bar{x}) \notin D_j^{(t+1)}, (x,\underline{x}) \notin D_j^{(t+1)}, 1 \leqslant j \leqslant n \end{cases} \tag{2}$$

可见, 决策信息系统在动态环境中, 伴随着数据对象在内存窗口的移入移出变化, 必然会导致相关条件分类与决策分类发生相应变化, 其条件概率及三支区域必然也会发生变化。本文首先系统地讨论在线数据的动态变化情况下条件概率及三支区域的变化, 并以此为基础提出三支决策在线快速计算算法。

2　三支决策在线快速计算方法

上述模型实现了动态流数据在线计算的动态变化机制。在经典计算中, 对于信息系统的每一次数据更新, 采用等价类的重新划分与计算, 开销较大。基于内存滑动窗口的三支决策动态变化情况需要借鉴增量学习的思想, 对于移入移出数据进行同步处理, 从而能够加快三支决策在线计算状态下条件概率与三支区域的更新速度, 保证了动态三支决策更新的实时性与高效性。

2.1　条件概率快速估计方法

如何利用已有的决策知识, 实现条件概率的动态估计是基于三支决策知识更新的基础。

由定义 2 可知, 计算概率粗糙集三支决策的原理是: 首先计算每条决策规则的条件概率, 然后将每条决策规则的条件概率与三支决策阈值 α、β 进行比较, 然后才能确定每个条件等价类属于哪个三支决策区域。可见, 每条决策规则的条件概率计算, 是概率粗糙集三支决策的关键步骤。

定理 1　给定 t 时刻信息系统 $\mathrm{IS}^{(t)} = \{U^{(t)}, C^{(t)} \cup D^{(t)}\}$, t+1 时刻信息系统更新为 $\mathrm{IS}^{(t+1)} = \{U^{(t+1)}, C^{(t+1)} \cup D^{(t+1)}\}$, 在此期间, 基于内存滑动窗口机制, 移入单数据对象 $\bar{x}$, 同时移出单数据对象 $\underline{x}$, 当动态数据对象符合以下变化模式时, 条件概率呈增大趋势, 即 $\Pr(D_j^{(t+1)} \mid R_i^{(t+1)}) > \Pr(D_j^{(t)} \mid R_i^{(t)})$。

变化模式 1: $(\bar{x} \notin R_i^{(t+1)} \wedge \bar{x} \notin D_j^{(t+1)}) \wedge (\underline{x} \in R_i^{(t)} \wedge \underline{x} \notin D_j^{(t)})$

变化模式 2: $(\bar{x} \notin R_i^{(t+1)} \wedge \bar{x} \in D_j^{(t+1)}) \wedge (\underline{x} \in R_i^{(t)} \wedge \underline{x} \notin D_j^{(t)})$

变化模式 3: $(\bar{x} \in R_i^{(t+1)} \wedge \bar{x} \in D_j^{(t+1)}) \wedge \begin{cases} (\underline{x} \notin R_i^{(t)} \wedge \underline{x} \notin D_j^{(t)}) \\ (\underline{x} \in R_i^{(t)} \wedge \underline{x} \notin D_j^{(t)}) \\ (\underline{x} \notin R_i^{(t)} \wedge \underline{x} \in D_j^{(t)}) \end{cases}$

证明　若移入对象 $\bar{x}$ 及移出对象 $\underline{x}$ 满足变化模式 1, 则有 $R_i^{(t+1)} = R_i^{(t)} - \{\underline{x}\}, D_j^{(t+1)} = D_j^{(t)}$。根据定义 1 及给定 $t+1$ 时刻决策表 $\mathrm{IS}^{(t+1)}$ 的数据变化公式 (1) 和 (2), 可推断出 $t+1$ 时刻条件概率 $\Pr(D_j^{(t+1)} \mid R_i^{(t+1)})$ 更新为

$$\Pr(D_j^{(t+1)} \mid R_i^{(t+1)}) = \frac{\left|D_j^{(t+1)} \cap R_i^{(t+1)}\right|}{|R_i^{(t+1)}|} = \frac{\left|(D_j^{(t)}) \cap (R_i^{(t)} - x)\right|}{\left|R_i^{(t)} - \{\underline{x}\}\right|} = \frac{\left|(D_j^{(t)}) \cap (R_i^{(t)})\right|}{|R_i^{(t)}| - 1} > \frac{\left|D_j^{(t)} \cap R_i^{(t)}\right|}{|R_i^{(t)}|} = \Pr(D_j^{(t)} \mid R_i^{(t)})$$

其余变化模式的证明过程与变化模式 1 类似, 在此省略。

定理 2　给定 t 时刻信息系统 $\mathrm{IS}^{(t)} = \{U^{(t)}, C^{(t)} \cup D^{(t)}\}$, t+1 时刻信息系统更新为 $\mathrm{IS}^{(t+1)} = \{U^{(t+1)}, C^{(t+1)} \cup D^{(t+1)}\}$, 在此期间, 基于动态流数据滑动窗口机制, 移入单数据对象 $\bar{x}$, 同时移出单数据对象 $\underline{x}$, 当流数据对象符合以下变化模式时, 条件概率呈减小趋势, 即 $\Pr(D_j^{(t+1)} \mid R_i^{(t+1)}) < \Pr(D_j^{(t)} \mid R_i^{(t)})$。

变化模式 4: $(\overline{x}\notin R_i^{(t+1)}\wedge\overline{x}\notin D_j^{(t+1)})\wedge(\underline{x}\in R_i^{(t)}\wedge\underline{x}\in D_j^{(t)})$

变化模式 5: $(\overline{x}\notin R_i^{(t+1)}\wedge\overline{x}\in D_j^{(t+1)})\wedge(\underline{x}\in R_i^{(t)}\wedge\underline{x}\in D_j^{(t)})$

变化模式 6: $(\overline{x}\in R_i^{(t+1)}\wedge\overline{x}\notin D_j^{(t+1)})\wedge\begin{cases}(\underline{x}\notin R_i^{(t)}\wedge\underline{x}\notin D_j^{(t)})\\(\underline{x}\notin R_i^{(t)}\wedge\underline{x}\in D_j^{(t)})\\(\underline{x}\in R_i^{(t)}\wedge\underline{x}\in D_j^{(t)})\end{cases}$

证明 若移入对象$\overline{x}$及移出对象$\underline{x}$遵循变化模式 4, 则有$R_i^{(t+1)}=R_i^{(t)}-\{\underline{x}\},D_j^{(t+1)}=D_j^{(t)}-\{\underline{x}\}$。根据定义 1 及给定$t+1$时刻决策表$\mathrm{IS}^{(t+1)}$的数据变化公式 (1)、(2), 可推断出$t+1$时刻条件概率$\Pr(D_j^{(t+1)}|R_i^{(t+1)})$更新为

$$\Pr(D_j^{(t+1)}\left|R_i^{(t+1)}\right.)=\frac{\left|D_j^{(t+1)}\cap R_i^{(t+1)}\right|}{|R_i^{(t+1)}|}=\frac{\left|(D_j^{(t)}-\{\underline{x}\})\cap(R_i^{(t)}-\{\underline{x}\})\right|}{\left|R_i^{(t)}-\{\underline{x}\}\right|}=\frac{\left|D_j^{(t)}\cap R_i^{(t)}\right|-1}{|R_i^{(t)}|-1}<\frac{\left|D_j^{(t)}\cap R_i^{(t)}\right|}{|R_i^{(t)}|}=\Pr(D_j^{(t)}\left|R_i^{(t)}\right.)$$

其余变化模式的证明过程与变化模式 4 类似, 在此省略。

定理 3 给定t时刻信息系统$\mathrm{IS}^{(t)}=\{U^{(t)},C^{(t)}\cup D^{(t)}\}$, $t+1$ 时刻信息系统更新为$\mathrm{IS}^{(t+1)}=\{U^{(t+1)},C^{(t+1)}\cup D^{(t+1)}\}$, 在此期间, 基于动态流数据滑动窗口机制, 移入单数据对象$\overline{x}$, 同时移出单数据对象$\underline{x}$, 当流数据对象符合变化模式 5~8 时, 条件概率保持不变, 即$\Pr(D_j^{(t+1)}\left|R_i^{(t+1)}\right.)=\Pr(D_j^{(t)}\left|R_i^{(t)}\right.)$。

变化模式 5: $(\overline{x}\notin R_i^{(t+1)}\wedge\overline{x}\notin D_j^{(t+1)})\wedge\begin{cases}(\underline{x}\notin R_i^{(t)}\wedge\underline{x}\notin D_j^{(t)})\\(\underline{x}\notin R_i^{(t)}\wedge\underline{x}\in D_j^{(t)})\end{cases}$

变化模式 6: $(\overline{x}\notin R_i^{(t+1)}\wedge\overline{x}\in D_j^{(t+1)})\wedge\begin{cases}(\underline{x}\notin R_i^{(t)}\wedge\underline{x}\notin D_j^{(t)})\\(\underline{x}\notin R_i^{(t)}\wedge\underline{x}\in D_j^{(t)})\end{cases}$

变化模式 7: $(\overline{x}\in R_i^{(t+1)}\wedge\overline{x}\notin D_j^{(t+1)})\wedge(\underline{x}\in R_i^{(t)}\wedge\underline{x}\notin D_j^{(t)})$

变化模式 8: $(\overline{x}\in R_i^{(t+1)}\wedge\overline{x}\in D_j^{(t+1)})\wedge(\underline{x}\in R_i^{(t)}\wedge\underline{x}\in D_j^{(t)})$

证明 若移入对象$\overline{x}$及移出对象$\underline{x}$遵循变化模式 5: $(\overline{x}\notin R_i^{(t+1)}\wedge\overline{x}\notin D_j^{(t+1)})\wedge(\underline{x}\notin R_i^{(t)}\wedge\underline{x}\notin D_j^{(t)})$, 则有$R_i^{(t+1)}=R_i^{(t)}\wedge D_j^{(t+1)}=D_j^{(t)}$。根据定义 1 及给定$t+1$时刻决策表$\mathrm{IS}^{(t+1)}$的数据变化公式 (1) 和公式 (2), 可推断出$t+1$时刻条件概率$\Pr(D_j^{(t+1)}\left|R_i^{(t+1)}\right.)$更新为

$$\Pr(D_j^{(t+1)}\left|R_i^{(t+1)}\right.)=\frac{|D_j^{(t+1)}\cap R_i^{(t+1)}|}{|R_i^{(t+1)}|}=\frac{|(D_j^{(t)}\cap R_i^{(t)})|}{|R_i^{(t)}|}=\Pr(D_j^{(t)}\left|R_i^{(t)}\right.)$$

其余变化模式的证明过程与变化模式 5 类似, 在此省略。通过定理 1~3 可知, 随着决策系统中数据对象同步移入移出的更新, 原有决策规则的条件概率的变化趋势可以通过局部更新数据的计算进行快速估计。这样既避免了原有数据对象的重复学习, 又利用同步更新大大提高了条件概率的计算效率, 实现了三支决策条件概率的在线计算, 进而使得三支决策区域的在线更新成为可能。

2.2 三支决策区域在线更新方法

三支决策的在线快速计算, 即充分利用t时刻信息系统的获取的三支决策知识以及$t+1$时刻变化数据的动态变化信息, 快速且准确地计算三支决策的区域变化。其计算的原理是根据 2.1 节定理 1~3 获得的各个决策规则的条件概率变化趋势及特定条件概率取值, 推导出三支决策区域在$t+1$时刻的变化规律。

推论 1 给定t时刻信息系统$\mathrm{IS}^{(t)}=\{U^{(t)},C^{(t)}\cup D^{(t)}\}$和$t+1$时刻信息系统$\mathrm{IS}^{(t+1)}=\{U^{(t+1)},C^{(t+1)}\cup D^{(t+1)}\}$, 若定理 1 描述的变化模式成立, 三支决策区域更新如下:

$$1)\ \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})=\begin{cases}\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})-R_i^{(t)}\cup R_i^{(t+1)}, & p_1\\ \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})\cup R_i^{(t+1)}, & p_2\\ \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})\cup R_i^{(t+1)}, & p_3\end{cases}$$

其中

$$p_1: R_i^{(t)}\subseteq\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})$$
$$p_2: R_i^{(t)}\subseteq\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})\wedge\Pr(D_j^{(t+1)}|R_i^{(t+1)})\geqslant\alpha$$
$$p_3: R_i^{(t)}\subseteq\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})\wedge\Pr(D_j^{(t+1)}|R_i^{(t+1)})\geqslant\alpha$$

$$2)\ \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)})=\begin{cases}\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})-R_i^{(t)}, & b_1\\ \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})-R_i^{(t)}\cup R_i^{(t+1)}, & b_2\\ \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})\cup R_i^{(t+1)}, & b_3\end{cases}$$

其中

$$b_1: R_i^{(t)}\subseteq\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})\wedge\Pr(D_j^{(t+1)}|R_i^{(t+1)})\geqslant\alpha$$
$$b_2: R_i^{(t)}\subseteq\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})\wedge\beta<\Pr(D_j^{(t+1)}|R_i^{(t+1)})<\alpha$$
$$b_3: R_i^{(t)}\subseteq\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})\wedge\beta<\Pr(D_j^{(t+1)}|R_i^{(t+1)})<\alpha$$

$$3)\ \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)})=\begin{cases}\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})-R_i^{(t)}, & n_1\\ \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})-R_i^{(t+1)}, & n_2\\ \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})-R_i^{(t)}\cup R_i^{(t+1)}, & n_3\end{cases}$$

其中

$$n_1: R_i^{(t)}\subseteq\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})\wedge\Pr(D_j^{(t+1)}|R_i^{(t+1)})\geqslant\alpha$$
$$n_2: R_i^{(t)}\subseteq\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})\wedge\beta<\Pr(D_j^{(t+1)}|R_i^{(t+1)})<\alpha$$
$$n_3: R_i^{(t)}\subseteq\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})\wedge\Pr(D_j^{(t+1)}|R_i^{(t+1)})\leqslant\beta$$

证明 1) 对于$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$

①假设$p_1: R_i^{(t)}\subseteq\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})$成立, 由定义 2 得$\Pr(D_j^{(t)}|R_i^{(t)})>\alpha$。由定理 1 可知$\Pr(D_j^{(t+1)}|R_i^{(t+1)})>\Pr(D_j^{(t)}|R_i^{(t)})>\alpha$, 即$R_i^{(t+1)}\subseteq\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$, 则接受域更新为$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})=\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})-R_i^{(t)}\cup R_i^{(t+1)}$。

②假设$p_2: R_i^{(t)}\subseteq\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})\wedge\Pr(D_j^{(t+1)}|R_i^{(t+1)})\geqslant\alpha$成立, 由定义 2 得$\beta<\Pr(D_j^{(t)}|R_i^{(t)})<\alpha$。因为$\Pr(D_j^{(t+1)}|R_i^{(t+1)})\geqslant\alpha$, 即$R_i^{(t+1)}\subseteq\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$, 则接受域更新为$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})=\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})\cup R_i^{(t+1)}$。

③假设 $p_3: R_i^{(t)} \subseteq \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \geqslant \alpha$ 成立, 由定义 2 得$\Pr(D_j^{(t)}|R_i^{(t)}) \leqslant \beta$。因为$\Pr(D_j^{(t+1)}|R_i^{(t+1)}) \geqslant \alpha$, 即$R_i^{(t+1)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$, 则接受域更新为$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) \cup R_i^{(t+1)}$。

2) 对于$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)})$

①假设$b_1: R_i^{(t)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \geqslant \alpha$成立, 由定义 2 得$\beta < \Pr(D_j^{(t)}|R_i^{(t)}) < \alpha$。因为$\Pr(D_j^{(t+1)}|R_i^{(t+1)}) \geqslant \alpha$, 即$R_i^{(t+1)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})$, 则边界域更新为$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)}$。

②假设$b_2: R_i^{(t)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$成立, 由定义 2 得$\beta < \Pr(D_j^{(t)}|R_i^{(t)}) < \alpha$, 因为$\beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$, 即$R_i^{(t+1)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})$, 则边界域更新为$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)} \cup R_i^{(t+1)}$。

③假设$b_3: R_i^{(t)} \subseteq \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$成立, 由定义 2 得$\Pr(D_j^{(t)}|R_i^{(t)}) < \beta$, 因为$\beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$, 即$R_i^{(t+1)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})$, 则边界域更新为$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \cup R_i^{(t+1)}$。

3) 对于$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)})$

①假设$n_1: R_i^{(t)} \subseteq \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \geqslant \alpha$成立, 由定义 2 得$\Pr(D_j^{(t)}|R_i^{(t)}) \geqslant \alpha$, 因为$\Pr(D_j^{(t+1)}|R_i^{(t+1)}) \geqslant \alpha$, 即$R_i^{(t+1)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})$, 则拒绝域更新为$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)}$。

②假设$n_2: R_i^{(t)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$成立, 由定义 2 得$\beta < \Pr(D_j^{(t)}|R_i^{(t)}) < \alpha$, 因为$\beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$, 即$R_i^{(t+1)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)})$, 则拒绝域更新为$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t+1)}$。

③假设$n_3: R_i^{(t)} \subseteq \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \leqslant \beta$成立, 由定义 2 得$\Pr(D_j^{(t)}|R_i^{(t)}) < \beta$, 因为$\Pr(D_j^{(t+1)}|R_i^{(t+1)}) \leqslant \beta$, 即$R_i^{(t+1)} \subseteq \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)})$, 则拒绝域更新为$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)} \cup R_i^{(t+1)}$。

证毕。

推论 2　给定 t 时刻信息系统$\mathrm{IS}^{(t)} = \{U^{(t)}, C^{(t)} \cup D^{(t)}\}$和 $t+1$ 时刻信息系统$\mathrm{IS}^{(t+1)} = \{U^{(t+1)}, C^{(t+1)} \cup D^{(t+1)}\}$, 若定理 2 描述的变化模式成立, 则三支决策区域更新为

$$1)\ \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)}) = \begin{cases} \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)} \cup R_i^{(t+1)}, & p_1 \\ \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)}, & p_2 \\ \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)}, & p_3 \end{cases}$$

式中:

$$p_1: R_i^{(t)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \geqslant \alpha$$
$$p_2: R_i^{(t)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$$
$$p_3: R_i^{(t)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \leqslant \beta$$

$$2)\ \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)}) = \begin{cases} \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \cup R_i^{(t+1)}, & b_1 \\ \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)} \cup R_i^{(t+1)}, & b_2 \\ \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)}, & b_3 \end{cases}$$

式中:

$$b_1: R_i^{(t)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$$
$$b_2: R_i^{(t)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \beta < \Pr(D_j^{(t+1)}|R_i^{(t+1)}) < \alpha$$
$$b_3: R_i^{(t)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \leqslant \beta$$

$$3)\ \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)}) = \begin{cases} \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) \cup R_i^{(t+1)}, & n_1 \\ \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) \cup R_i^{(t+1)}, & n_2 \\ \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) - R_i^{(t)} \cup R_i^{(t+1)}, & n_3 \end{cases}$$

式中:

$$n_1: R_i^{(t)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \leqslant \beta$$
$$n_2: R_i^{(t)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \wedge \Pr(D_j^{(t+1)}|R_i^{(t+1)}) \leqslant \beta$$
$$n_3: R_i^{(t)} \subseteq \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})$$

证明　证明过程与推论 1 类似, 此处略。证毕。

推论 3　给定 t 时刻信息系统$\mathrm{IS}^{(t)} = \{U^{(t)}, C^{(t)} \cup D^{(t)}\}$和 $t+1$ 时刻信息系统$\mathrm{IS}^{(t+1)} = \{U^{(t+1)}, C^{(t+1)} \cup D^{(t+1)}\}$, 若定理 3 描述的变化模式成立, 三支决策区域更新为:

$$\mathrm{POS}: R_i^{(t)} \subseteq \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)}) \Rightarrow \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)});$$
$$\mathrm{BND}: R_i^{(t)} \subseteq \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)}) \Rightarrow \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)});$$
$$\mathrm{NEG}: R_i^{(t)} \subseteq \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)}) \Rightarrow \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)}) = \mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})。$$

证明　根据定理 3, 条件概率$\Pr(D_j^{(t+1)} \big| C_i^{(t+1)})$保持不变, 由定义 2 可知此时三支决策区域亦是保持不变的。证毕。

2.3　三支决策在线快速计算算法

根据 2.1 节中的定理 1~3 推导出的概率粗糙集的条件概率变化趋势, 以及 2.2 节中的三支决策区域的快速计算 3 种情形展开讨论。

本节设计了三支决策在线快速计算算法 (online computing algorithm), 并从算法时间复杂度角度与基于三支决策理论的经典计算算法作对比, 从理论上分析在线快速计算算法的优势。

算法 1　三支决策在线快速计算算法

输入　t 时刻信息系统$\mathrm{IS}^{(t)} = \{U^{(t)}, C^{(t)} \cup D^{(t)}\}$, 条件等价类和决策等价类$R_i^{(t)}$、$D_j^{(t)}(i = 1,2,\cdots,m; j = 1,2,\cdots,n)$, 条件概率$\Pr(D_j^{(t)}|R_i^{(t)})(i = 1,2,\cdots,m; j = 1,2,\cdots,n)$, 三支决策接受域、拒绝域和边界域$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})$、$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})$及$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})$, 新增决策对象$\bar{x}$及被移出对象$\underline{x}$, 三支决策阈值 (α, β)。

输出　$t+1$ 时刻三支决策区域$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$、$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)})$及$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)})$。

1) 通过$\bar{x}$及$\underline{x}$的等价类从属关系, 更新 $t+1$ 时刻的条件等价类和决策等价类$R_i^{(t+1)}$和$D_j^{(t+1)}$;

2) 根据定理 1~3 评估 $t+1$ 时刻条件概率$\Pr(D_j^{(t+1)}|R_i^{(t+1)})$的变化趋势;

3) 根据推论 1~3 更新 $t+1$ 时刻三支决策区域$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$、$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)})$、$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)})$;

4) 更新 t+1 时刻每个等价类的条件概率值 $\Pr(D_j^{(t+1)}|R_i^{(t+1)})$及条件、决策等价类$R^{(t+1)}$、$D^{(t+1)}$, 返回 1)。

在线快速计算中, 若给定内存部分信息系统 IS, 包含 m 个条件等价类和 n 个决策等价类, 其中 $m=|U^{(t)}/A|(U^{(t)}\subseteq U)$, $n=|D|$, 等价类的计算时间复杂度为$O(|U^{(t)}/A|\cdot|A|+|D|)$。步骤 2) 评估条件概率变化趋势, 其时间复杂度为$O(1)$, 这时可以忽略不计的。步骤 3) 为更新三支区域, 即区域内的对象可能发生转移, 区域未发生变化是最好的情况, 在这种情况下两种算法的时间复杂度均可视为$O(1)$; 最坏的情况, 此算法中需要转移对象两次, 以$|R_i|$和$|R_j|$表示移入对象$\bar{x}$和被移出对象$\underline{x}$的条件和决策等价类, 则区域更新的时间复杂度为$O(|R_i|+|R_j|)$。步骤 4) 中, 在$|R_i|$和$|D_j|$已知的情况下, 重新计算条件概率的时间复杂度几乎是可以忽略不计的。综上, 在线快速计算的时间复杂度为$O(|U^{(t)}/A|\cdot|A|+|D|)+O(|R_i|+|R_j|)$。

为了说明上述在线三支决策算法的特点与优势, 本文将根据文献[22]中提出的基于概率粗糙集三支决策理论的经典计算思想, 设计三支决策经典计算算法 (classical computing algorithm)。

算法 2　基于三支决策的经典计算算法

输入　t 时刻下信息系统$\mathrm{IS}^{(t)}=\{U^{(t)},C^{(t)}\cup D^{(t)}\}$, 条件等价类和决策等价类$R_i^{(t)}$、$D_j^{(t)}(i=1,2,\cdots,m;j=1,2,\cdots,n)$, 条件概率$\Pr(D_j^{(t)}|R_i^{(t)})(i=1,2,\cdots,m;j=1,2,\cdots,n)$, 三支决策接受域、拒绝域和边界域$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t)})$、$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t)})$、$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t)})$, 新增决策对象$\bar{x}$及被移出对象$\underline{x}$, 三支决策阈值 (α,β)。

输出　t+1 时刻的三支决策区域及决策规则$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$、$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)})$和$\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)})$。

/* 根据移入对象$\bar{x}$、移出对象$\underline{x}$进行三支决策更新计算*/

1) 通过$\bar{x}$和$\underline{x}$的等价类从属关系, 更新 t+1 时刻的条件等价类和决策等价类$R_i^{(t+1)}$和$D_j^{(t+1)}$;

2) 根据定义 1 重新计算 t+1 时刻条件概率$\Pr(D_j^{(t+1)}|R_i^{(t+1)})$值;

3) 根据定义 2 重新计算 t+1 时刻三支决策接受域、拒绝域和边界域$\mathrm{POS}_{(\alpha,\beta)}(D_j^{(t+1)})$、$\mathrm{BND}_{(\alpha,\beta)}(D_j^{(t+1)})$, $\mathrm{NEG}_{(\alpha,\beta)}(D_j^{(t+1)})$。

经典计算算法中, 若给定内存部分信息系统 IS, 包含 m 个条件等价类和 n 个决策等价类, 其中 $m=|U^{(t)}/A|(U^{(t)}\subseteq U)$, $n=|D|$, 等价类的计算时间复杂度为$O(|U^{(t)}/A|\cdot|A|+|D|)$; 重新计算条件概率的时间复杂度为$O(|U^{(t)}/A|\cdot|D|)$; 根据条件概率大小与阈值的比较, 更新三支区域的时间复杂度亦为$O(|U^{(t)}/A|\cdot|D|)$。因此, 经典计算算法的时间复杂度为$O(|U^{(t)}/A|\cdot|A|+|D|)+O(2|U^{(t)}/A|\cdot|D|)$。

假设原 t 时刻信息系统为$\mathrm{IS}^{(t)}$, 新增对象集为$\{\bar{x}\}$, 上述两种算法的时间复杂度分别为:

经典计算算法:

$$O(|U^{(t)}/A|\cdot|A|+|D|)+O(2|U^{(t)}/A|\cdot|D|)$$

在线快速计算:

$$O(|U^{(t)}/A|\cdot|A|+|D|)+O(|R_i|+|R_j|)$$

从时间复杂度分析可知, 在线快速计算的效率明显优于经典计算算法。从算法执行过程亦可看出, 由于经典计算算法对于决策对象的每一次更新, 都进行等价类划分和条件概率的重新计算, 其开销甚是庞大; 而在线计算利用内存滑动窗口模型, 同步处理移入移出数据, 对当前实时变化的数据进行局部计算, 实现条件概率及三支区域的快速估计与更新, 其运算效率明显高于经典计算算法。

综上, 从理论分析可知在线快速计算的运行效率明显优于经典计算算法。接下来, 我们将从实验的角度评估两算法的实际表现。

3　实验与结果

2.3 节通过理论证明了三支决策在线快速计算算法的时间复杂度优于经典计算算法。本章将通过与三支决策经典计算算法的对比实验来验证三支决策在线快速计算算法的时间消耗优势。

本实验环境为: 双核、4 GB 内存的 PC, 运行 Microsoft Windows 7 操作系统, 算法程序使用 Java 编程, Java 开发包为 JDK1.6 版本。

3.1　数据集

为了验证在线快速计算的性能, 我们从 UCI 数据库中选择了 8 个较大数据集, 分别为 Skin-NoSkin、Shuttle、IRIS、Zoo、Haberman、Breast-cancer、Letter、Magic, 所有的数据均为数值型或类别型, 表 1 为数据集的基本信息。

表 1　数据集基本信息

Table 1　The basic information of the eight datasets

序号	数据集	样本数	特征数	概念计数
1	Skin_NoSkin	64 486	3	2
2	Shuttle	58 000	9	5
3	IRIS	60 750	8	3
4	Zoo	60 600	17	7
5	Haberman	55 080	3	2
6	Breast-cancer	60 860	10	4
7	Letter	20 000	16	12
8	Magic	19 020	10	2

3.2　实验和结果

为排除偶然因素的影响, 所有实验都进行了 10 次, 最终结果是 10 次实验的平均值。我们任意设定的三支决策阈值(α,β)为 (0.75, 0.3)。

1) 算法执行时间对比实验

假设存储空间保持为 100 条记录, 我们比较在线快速计算、经典计算算法的性能时, 逐渐增加上述 8 个数据集在内存中存储数据的比例。当进入决策信息系统的在线计算数据总量累计达到总数据量的 10%~100% 时, 分别测试不同情况下, 两种算法在三支决策区域更新中的平均执行时间。实验结果如图 2 所示。

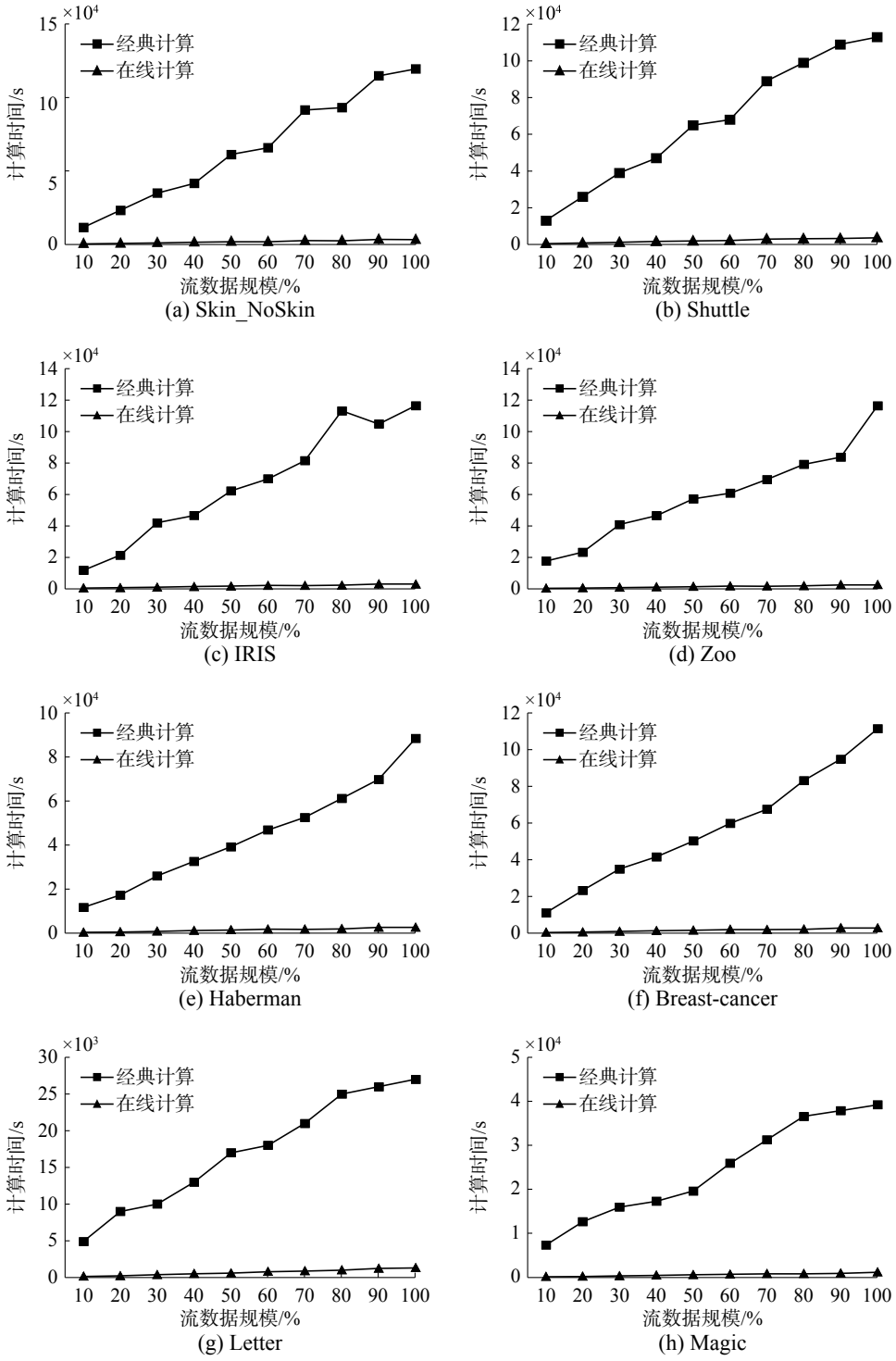


图 2　在线快速计算与经典计算算法的平均时间消耗对比

Fig. 2　Average elapsed times between online computing algorithm and classical computing algorithm

由图 2 显然可以看出, 在每个数据集下, 随着进入内存中动态数据比例的增大, 在线快速计算与经典计算算法的平均执行时间亦逐渐增大, 但经典计算算法的增幅明显更大。可见, 在线快速

计算的平均执行时间相对更少。

此外，为了更加精确地展现在线快速计算的优势及其稳定性，我们将在占数据总量 50% 的情况下，从 t 时刻到 t+5 时刻，测试两类算法的平均执行时间并对比其稳定性。结果见表 2，单位为 ms。可以观察到，对于同一数据集，对比经典计算算法，在线快速计算可以大大节约算法执行时间。

表 2　t ~ t+5 时刻算法执行时间

Table 2　Experimental results examined from time t to t+5 on UCI datasets　　ms

数据集	算法	t	t+1	t+2	t+3	t+4	t+5	平均值±标准差
Skin_NoSkin	经典计算	2.280	1.860	1.890	1.770	1.740	1.800	1.890±0.007
	在线计算	0.055	0.043	0.040	0.039	0.039	0.038	0.042±0.006
Shuttle	经典计算	2.220	2.040	1.890	1.860	1.980	1.860	1.980±0.005
	在线计算	0.063	0.046	0.045	0.043	0.041	0.039	0.046±0.009
IRIS	经典计算	2.370	2.280	2.070	1.860	1.980	1.860	1.980±0.005
	在线计算	0.058	0.048	0.040	0.041	0.040	0.040	0.044±0.007
Zoo	经典计算	2.340	1.920	2.100	1.800	1.770	1.800	1.950±0.008
	在线计算	0.050	0.040	0.037	0.039	0.037	0.036	0.040±0.005
Haberman	经典计算	1.950	1.620	1.500	1.500	1.440	1.500	1.590±0.006
	在线计算	0.043	0.034	0.038	0.031	0.030	0.029	0.034±0.005
Breast-cancer	经典计算	1.890	1.710	1.710	1.650	1.680	1.740	1.740±0.003
	在线计算	0.044	0.036	0.034	0.033	0.031	0.032	0.035±0.005
Letter	经典计算	2.200	0.740	1.500	1.420	1.400	1.440	1.620±0.015
	在线计算	0.079	0.074	0.058	0.052	0.051	0.049	0.061±0.013
Magic	经典计算	2.200	1.860	1.560	1.440	1.320	1.340	1.620±0.017
	在线计算	0.071	0.059	0.053	0.041	0.039	0.043	0.053±0.012

同时，在不同数据集上，在线快速计算执行时间的均值和方差比经典计算算法更小，显示了其具有更好的效率及稳定性。

2) 不同内存容量算法执行时间对比

由于在线计算方法以内存计算为依托，内存空间的容量对于三支决策在线快速计算的执行效果是否有影响，是一个值得研究的问题。因此我们选择以上 8 个 UCI 数据集中规模较大且复杂程度较高的 Breast-cancer 数据集，构造和实验 1 相似的对比实验，但是将内存容量分别为设定为 100、500、1 000、2 000 条记录，当上述数据集实施两种算法实验的数据总量累计达到总数据量的 10%、40%、70% 和 100% 时，分别记录这两种算法此时的平均执行时间。

由图 3 可以观察到，在相同实验数据累积量下，随着内存空间的增加，经典计算算法的平均执行时间几乎呈指数级速度增长，而在线快速计算的时间消耗较为平稳。

由算法的时间复杂度可知，随着决策系统的每一次更新，经典计算算法需要进行所有数据等价类的重新划分及条件概率的重新计算，此过程每次都需要执行时间复杂度为$O(2|U^{(t)}/A|\cdot|D|)$的步骤，其时间消耗极大，且与内存中的数据量及复杂程度密切相关；而在线快速计算由于只需要对当前实时变化的数据对象进行局部计算，避免了历史数据的重复学习，最多只需执行时间复杂度为$O(\mathrm{U}^{(t)}/\mathrm{A}|\cdot|\mathrm{D}|)$的步骤，大大节省了此更新过程的开销。

综上可知，在不同内存空间下，在线快速计算的效率均远高于经典计算算法，且这一优势随着内存空间的增加更加明显。

3) 不同阈值在线快速计算执行时间对比

由于三支决策是一种典型的概率粗糙集参数化模型，实验结果可能会受参数设置的影响。设置内存空间为 100 b，并选取了 5 对阈值，分别测出在线快速计算在 8 个数据集下执行完毕平均所需时间，从而验证在线快速计算的性能与选取阈值的关系。图 4 显示了算法平均执行时间随不同阈值对的变化趋势，不难看出，算法的执行时间随着阈值的改变有轻微的波动，总体来说比较稳定。由此也进一步验证了在线快速计算的时间复杂度主要取决于动态数据的规模及其复杂程度。

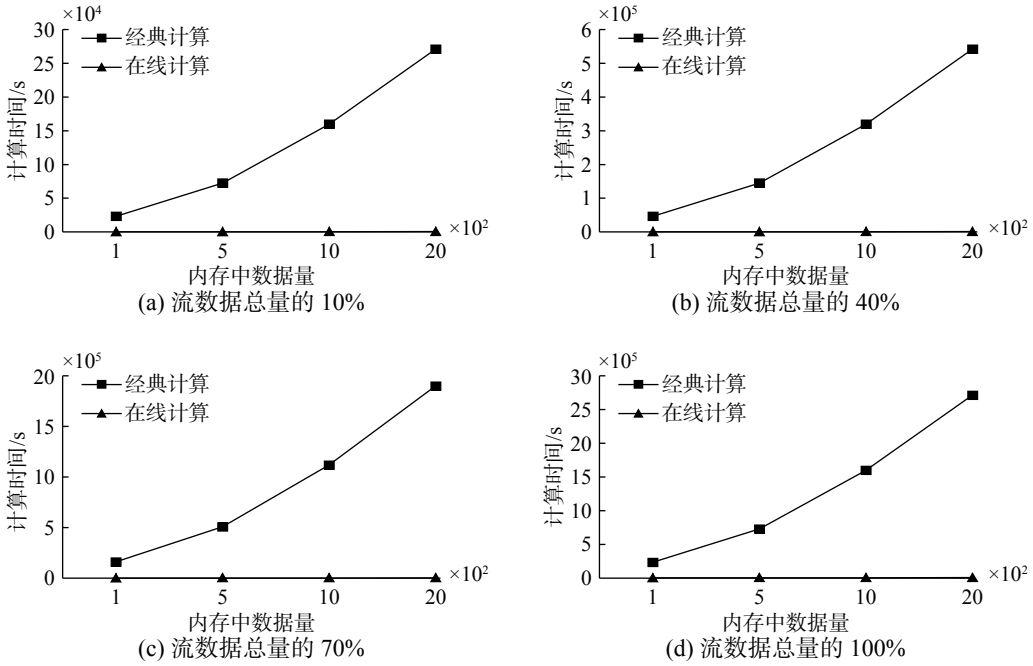


图 3　在线快速计算与经典计算算法在不同内存容量下的平均执行时间

Fig. 3　Average elapsed times between online computing algorithm and classical computing algorithm on different memory sizes

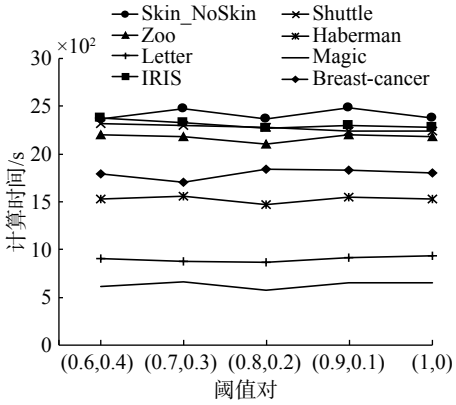


图 4　不同阈值对在线快速计算平均执行时间

Fig. 4　Average elapsed times on online computing algorithm with different threshold pairs

4　结束语

动态在线计算是一种常见的数据计算方法，本文提出一种内存滑动窗口机制对在线计算的数据变化模式进行统一建模，并根据上述模型，推导出不同类型数据变化模式下的三支决策条件概率及三支区域的变化规律。最后本研究提出一种新型的动态在线快速计算算法，该算法能够获得与经典三支决策算法相同的决策结果。由于本算法只需要对当前的变化数据进行局部实时计算，有效避免了历史数据的重复学习，大大提高了三支决策更新效率。为验证本文所提出的在线快速计算的优势，与经典计算算法在 8 个 UCI 公共数据集上进行实验对比。实验结果表明，在线快速计算比经典计算算法效率更高，稳定性更好。本研究表明，概率粗糙集三支决策领域采用在线计算方法进行快速计算是可行的。我们将在未来的工作中，继续探讨概率粗糙集三支决策的多对象动态在线快速计算以及其在实际在线计算场景中的应用。

参考文献:

[1] FONG S, WONG R, VASILAKOS A V. Accelerated PSO swarm search feature selection for data stream mining big data[J]. IEEE transactions on services computing, 2016, 9(1): 33–45.

[2] ZHANG Hong, LI Bo, JIANG Hongbo, et al. A framework for truthful online auctions in cloud computing with heterogeneous user demands[C]//Proceedings of 2013 IEEE INFOCOM. Turin, Italy, 2013: 1510–1518.

[3] ESKANDARI S, JAVIDI M M. Online streaming feature selection using rough sets[J]. International journal of approximate reasoning, 2016, 69: 35–57.

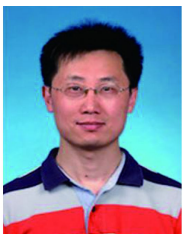
[4] CHAZELLE B, ROSENBERG B. The complexity of computing partial sums off-line[J]. International journal of computational geometry and applications, 1991, 1(1): 33–45.

[5] DENG Jie, QU Zhiguo, ZHU Yongxu, et al. Towards efficient and scalable data mining using spark[C]//Proceedings of 2014 International Conference on Information and Communications Technologies. Nanjing, China, 2014: 1–6.

[6] YAO Yiyu. Three-way decisions and cognitive computing [J]. Cognitive computation, 2016, 8(4): 543–554.

[7] ZHOU Bing, YAO Yiyu, LUO Jigang. Cost-sensitive three-way email spam filtering[J]. Journal of intelligent information systems, 2014, 42(1): 19–45.
[8] KHAN M T, AZAM N, KHALID S, et al. A three-way approach for learning rules in automatic knowledge-based topic models[J]. International journal of approximate reasoning, 2017, 82: 210–226.
[9] LI Huaxiong, ZHANG Libo, ZHOU Xianzhong, et al. Cost-sensitive sequential three-way decision modeling using a deep neural network[J]. International journal of approximate reasoning, 2017, 85: 68–78.
[10] QIAN Jin, DANG Chuangyin, YUE Xiaodong, et al. Attribute reduction for sequential three-way decisions under dynamic granulation[J]. International journal of approximate reasoning, 2017, 85: 196–216.
[11] YU Hong, ZHANG Cong, WANG Guoyin. A tree-based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge-based systems, 2016, 91: 189–203.
[12] LUO Chuan, LI Tianrui, CHEN Hongmei, et al. Efficient updating of probabilistic approximations with incremental objects[J]. Knowledge-based systems, 2016, 109: 71–83.
[13] NAUMAN M, AZAM N, YAO Jingtao. A three-way decision making approach to malware analysis using probabilistic rough sets[M]. New York, NY, USA: Elsevier, 2016.
[14] CHEN Yufei, YUE Xiaodong, FUJITA H, et al. Three-way decision support for diagnosis on focal liver lesions[J]. Knowledge-based systems, 2017, 127: 85–99.
[15] CHEN Hongmei, LI Tianrui, LUO Chuan, et al. A decision-theoretic rough set approach for dynamic data mining[J]. IEEE transactions on fuzzy systems, 2015, 23(6): 1958–1970.
[16] LIANG Decui, XU Zeshui, LIU Dun. Three-way decisions based on decision-theoretic rough sets with dual hesitant fuzzy information[J]. Information sciences, 2017, 396: 127–143.
[17] ZHANG Hongying, YANG Shuyun, MA Jianmin. Ranking interval sets based on inclusion measures and applications to three-way decisions[J]. Knowledge-based systems, 2016, 91: 62–70.
[18] ZHAO Xuerong, HU Baoqing. Fuzzy probabilistic rough sets and their corresponding three-way decisions[J]. Knowledge-based systems, 2016, 91: 126–142.
[19] AZAM N, ZHANG Yan, YAO Jingtao. Evaluation functions and decision conditions of three-way decisions with game-theoretic rough sets[J]. European journal of operational research, 2017, 261(2): 704–714.
[20] DENG Xiaofei, YAO Yiyu. A multifaceted analysis of probabilistic three-way decisions[J]. Fundamenta informaticae, 2014, 132(3): 291–313.
[21] ZHANG Yan, YAO Jingtao. Gini objective functions for three-way classifications[J]. International journal of approximate reasoning, 2017, 81: 103–114.
[22] GRECO S, MATARAZZO B, ROMAN S. Rough sets theory for multicriteria decision analysis[J]. European journal of operational research, 2001, 129(1): 1–47.
[23] LIU Dun, LI Tianrui, ZHANG Junbo. Incremental updating approximations in probabilistic rough sets under the variation of attributes[J]. Knowledge-based systems, 2015, 73: 81–96.
[24] LI Shaoyong, LI Tianrui, LIU Dun. Incremental updating approximations in dominance-based rough sets approach under the variation of the attribute set[J]. Knowledge-based systems, 2013, 40: 17–26.
[25] ZHANG Junbo, LI Tianrui, RUAN Da, et al. Neighborhood rough sets for dynamic data mining[J]. International journal of intelligent systems, 2012, 27(4): 317–342.
[26] LUO Chuan, LI Tianrui, CHEN Hongmei. Dynamic maintenance of three-way decision rules[C]//Proceedings of the 9th International Conference on Rough Sets and Knowledge Technology. Shanghai, China, 2014: 801–811.
[27] PAPANDREOU G, KOKKINOS I, SAVALLE P A. Modeling local and global deformations in Deep Learning: epitomic convolution, Multiple Instance Learning, and sliding window detection[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 2015: 390–399.
[28] YUN U, LEE G. Sliding window based weighted erasable stream pattern mining for stream data applications[J]. Future generation computer systems, 2016, 59: 1–20.

作者简介：



徐健锋，男，1973 年生，副教授，博士研究生，计算机学会会员，主要研究方向为数据挖掘、粗糙集、机器学习。主持国家自然基金项目 1 项，参与国家自然科学基金项目 2 项。

何宇凡，男，1994 年生，硕士研究生，主要研究方向为三支决策、粗糙集、粒计算、机器学习。

汤涛，男，1993 年生，硕士研究生，主要研究方向为粗糙集、粒计算、机器学习。