Computer Science and Application 计算机科学与应用, 2023, 13(12), 2364-2371
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312236

Hans汉斯

一种不完备决策系统下快速求取正域的算法

王　峰

烟台大学计算机与控制工程学院，山东　烟台

收稿日期：2023年11月25日；录用日期：2023年12月21日；发布日期：2023年12月27日

摘　要

现如今的互联网时代，数据维度灾难性增长，如何从高维数据中提取有用信息成为一大难题。由于数据的不完整性，不完备决策系统逐渐得到人们的关注。在数据中寻找确定性的规则是当今社会重要的研究方向之一，在不完备决策系统中正域的计算则代表着确定性的规则的提取。但效率一直是正域计算中的关键问题，正域的计算必须要通过求取相容类，计算相容类的复杂度直接影响到计算正域的复杂度。本文利用样本在属性集合下相容类的单调性给出了一个时间复杂度低的计算正域的算法。实验结果表明，本文方法在多个数据集上无论是维度还是规模上效率都是较高的，具有更好的稳定性，更适用于大规模以及大维度的数据。

关键词

粗糙集，不完备决策系统，正域，相容类

A Fast Algorithm for Computing Positive Regions in an Incomplete Decision System

Feng Wang

School of Computer and Control Engineering, Yantai University, Yantai Shandong

Received: Nov. 25th, 2023; accepted: Dec. 21st, 2023; published: Dec. 27th, 2023

Abstract

Nowadays, in the era of the Internet, data dimensions are growing catastrophically. Extracting useful information from high-dimensional data has become a major challenge. Due to the incompleteness of data, incomplete decision systems are gradually gaining attention. Finding deterministic rules in data is one of the important research directions in today's society, and computing

文章引用: 王峰. 一种不完备决策系统下快速求取正域的算法[J]. 计算机科学与应用, 2023, 13(12): 2364-2371.
DOI: 10.12677/csa.2023.1312236

the positive domain in incomplete decision systems represents the extraction of deterministic rules. However, efficiency has always been a key issue in the computation of positive domains. This computation involves finding tolerance classes, and the complexity of computing the tolerance classes directly affects the complexity of computing the positive domains. In this paper, we utilize the monotonicity of the tolerance classes of the samples under the set of attributes to propose an algorithm that computes the positive domain with low time complexity. Experimental results show that the method in this paper is efficient in both dimension and scale on multiple datasets, exhibits better stability, and is more suitable for large-scale as well as high-dimensional data.

Keywords

Rough Set, Incomplete Decision System, Positive Region, Tolerance Classes

1. 引言

在当代信息时代，数据的快速增长和复杂性给决策过程带来了巨大挑战。面对庞大、多样化的数据集，基于粗糙集理论[1]的传统的经典决策系统可能无法满足需求。因此，不完备决策系统[2] [3]成为了处理复杂决策问题的有效工具。不完备决策系统基于一种基本假设，即我们所拥有的信息并不完整，因此需要从有限的信息中进行决策。通过考虑不完整性和不确定性，不完备决策系统能够在实践中更好地适应真实世界的复杂性。

不完备决策系统的发展已经取得了显著的进展。研究者们提出了许多创新的方法和算法，用于有效地处理缺失数据[4] [5]、模糊信息[6] [7]、优势关系[8] [9]等问题。这些方法和算法使得不完备决策系统能够更好地应对现实中的不确定性，并提供可靠的决策支持。此外，随着机器学习、人工智能等领域的快速发展，不完备决策系统正逐渐融合这些技术，为决策制定者提供更准确、可靠的决策建议。

通过深入研究和分析，利用不完备决策系统解决现实世界中的复杂决策问题，寻找对应确定性的规则是一个重要的目标。在不完备决策系统中寻找确定性的规则，必须通过相容类的计算，而相容类的效率问题一直阻碍者确定性规则的提取速度。本文通过不完备决策系统的中容差类的单调性以及对正域的分析，提出了一种在不完备系统下快速寻找正域(确定性规则)的算法，经过实验证明本文所提算法的有效性以及稳定性。

2. 基本概念

不完备决策系统

由于数据采集或数据损坏等原因导致决策系统部分信息缺失，则称此类决策系统为不完备决策系统，本节将主要介绍不完备决策系统下的粗糙集理论相关概念，包括不完备决策系统、相容关系、上近似集、下近似集、正域、边界域和负域等相关概念。

定义 1 [10]在给定信息系统 $IS=(U,AT,V,f)$ 中，其中 $U=\{x_1,x_2,\cdots,x_{|U|}\}$ 表示非空有限对象的集合，称之为论域；AT 表示属性的非空有限集合；$V=\cup\{V_a|a\in AT\}$，其中 V_a 表示属性 $a\in AT$ 的值域；

$f:U\times AT\to V$ 是信息函数，$f(x,a)$ 表示对象 $x\in U$ 在属性 $a\in AT$ 上的值，简称 $a(x)$。

在信息系统 IS 中，若 $AT=C\cup D$，则该信息系统被称为决策系统，记作 $DS=(U,AT=C\cup D,V,f)$，其中 C 表示非空有限的条件属性的集合，D 表示非空有限的决策属性的集合。

定义 2 [10]在决策系统 $DS=(U,AT=C\cup D,V,f)$ 中，若存在 $a\in C$ 使得 V_a 含有缺失值，且缺失值使用 $*$ 表示，则该决策系统被称为不完备决策系统，记作 $IDS=(U,AT=C\cup D,V,f)$，其中决策属性 $d\in D$ 的值域 V_d 中均不含有缺失值。

定义 3 [10]在不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$ 下，对于 $\forall\mathbb{C}\in C$，存在相容关系：

$$TOL(\mathbb{C})=\left\{(x,y)\in U*U\,\middle|\,\forall a\in\mathbb{C},a(x)=a(y)\vee a(x)=*\vee a(y)=*\right\} \tag{1}$$

设 $\partial_{\mathbb{C}}(x)=\left\{y\in U\,\middle|\,(x,y)\in TOL(\mathbb{C})\right\}$，其中 $\partial_{\mathbb{C}}(x)$ 表示通过条件属性集合 $\mathbb{C}$ 与样本 x 不可区分的对象的集合，$\partial_{\mathbb{C}}(x)$ 也被称为相容类。因此，可以导出论域 U 在属性子集 $\mathbb{C}$ 上的覆盖 $U/TOL(\mathbb{C})=\left\{\partial_{\mathbb{C}}(x)\,\middle|\,x\in U\right\}$。

性质 1 不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$ 中，对于 $\forall\mathbb{C}\in C$，有：

$$TOL(\mathbb{C})=\bigcap_{a\in\mathbb{C}}TOL(a) \tag{2}$$

同时相容关系 $TOL(\mathbb{C})$ 满足：

1) 自反性：对 $\forall x\in U$，有 $(x,x)\in TOL(\mathbb{C})$。

2) 对称性：对 $\forall x,y\in U$，若 $(x,y)\in TOL(\mathbb{C})$，则 $(y,x)\in TOL(\mathbb{C})$。

定义 4 [11]在不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$ 中，对于 $\forall\mathbb{C}\subseteq C$，$\forall X\in U$，对于 X 关于 $\mathbb{C}$ 的下、上近似定义如下：

$$\underline{Appr_{\mathbb{C}}}(X)=\left\{x_i\in U\,\middle|\,\partial_{\mathbb{C}}(x_i)\subseteq X\right\} \tag{3}$$

$$\overline{Appr_{\mathbb{C}}}(X)=\left\{x_i\in U\,\middle|\,\partial_{\mathbb{C}}(x_i)\cap X\neq\varnothing\right\} \tag{4}$$

且正域表示为 $POS_{\mathbb{C}}(X)=\underline{Appr_{\mathbb{C}}}(X)$，边界域表示为 $BND_{\mathbb{C}}(X)=\overline{Appr_{\mathbb{C}}}(X)-\underline{Appr_{\mathbb{C}}}(X)$，负域为 $NEG_{\mathbb{C}}(X)=U-\overline{Appr_{\mathbb{C}}}(X)-\underline{Appr_{\mathbb{C}}}(X)$。

定义 5 [11]在不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$ 中，假设论域 U 对决策属性 D 上的划分为 $U/D=\{D_1,D_2,\cdots,D_m\}\ (1\le m\le|U|)$，对于 $\forall\mathbb{C}\subseteq C$，决策类集合 U/D 关于条件属性集合 $\mathbb{C}$ 的下、上近似定义如下：

$$\underline{Appr_{\mathbb{C}}}(U/D)=\left\{\underline{Appr_{\mathbb{C}}}(D_1),\underline{Appr_{\mathbb{C}}}(D_2),\cdots,\underline{Appr_{\mathbb{C}}}(D_m)\right\} \tag{5}$$

$$\overline{Appr_{\mathbb{C}}}(D)=\left(\overline{Appr_{\mathbb{C}}}(D_1),\overline{Appr_{\mathbb{C}}}(D_2),\cdots,\overline{Appr_{\mathbb{C}}}(D_m)\right) \tag{6}$$

决策类集合 U/D 关于条件属性 $\mathbb{C}\subseteq C$ 的正域以及边界域定义为：

$$POS_{\mathbb{C}}(U/D)=\bigcup_{D_i\in U/D}\underline{Appr_{\mathbb{C}}}(D_i) \tag{7}$$

$$BND_{\mathbb{C}}(U/D)=\bigcup_{D_i\in U/D}\overline{Appr_{\mathbb{C}}}(D_i)-\bigcup_{D_i\in U/D}\underline{Appr_{\mathbb{C}}}(D_i) \tag{8}$$

基于上述定理，表 1 给出不完备决策系统下的正域计算算法(Positive domain computational algorithms in incomplete decision systems, PDCA-IDS)。

Table 1. Positive domain computational algorithms in incomplete decision systems (PDCA-IDS)
表 1. 不完备决策系统下的正域计算算法(PDCA-IDS)

输入： 不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$。

输出： 正域集合 $POS_C(D)$。

1.初始化： $POS_C(D)=\varnothing$，对于 $\forall x_i\in U$， $\partial_C(x_i)=\varnothing$；

2.对于 $\forall x_i\in U$，计算 $\partial_C(x_i)$；

3.计算 $POS_C(D)$；

4.输出正域集合 $POS_C(D)$。

上述传统的计算正域的时间复杂度为 $O(mn^2)$，其中 m 为条件属性的个数(即 $|C|$)，n 为样本对象的个数。时间复杂度较高，效率低下。

3. 基于不完备决策系统的正域快速算法

由于数据量的增大，数据的维度和规模都是对不完备信息数据进行相容类的划分时的一个考验。常规的容差类的划分时间复杂度高、效率低下，本小节主要介绍相容类的单调性、对正域的分析以及正域快速算法的构建。

3.1. 相容类的快速计算

定理 1 [12]给定一个不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$，X_1 和 X_2 是 C 的两个属性子集，其中 $X_1\subseteq X_2\subseteq C$，对于 $\forall x_i\in U$， $\delta_{X_2}(x_i)\subseteq\delta_{X_1}(x_i)$。

定理 1 说明在当属性子集越大时，对任意样本对象 x_i 来说所含有的与其保持一致的样本对象越少，即对于任意的样本对象 x_i 在较大的属性子集下的相容类可以通过较小的属性子集的相容类得出。

Table 2. Fast positive domain computing algorithms for incomplete decision systems (FPDCA-IDS)
表 2. 不完备决策系统下快速正域计算算法(FPDCA-IDS)

输入： 不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$。

输出： 正域集合 $POS_C(D)$。

1.初始化： $POS_C(D)=\varnothing$，对于 $\forall x_i\in U$， $\partial_C(x_i)=\varnothing$；

2.任取 $a\in C$，对于 $\forall x_i\in U$ 计算 $\partial_a(x_i)$，计算 $POS_a(D)$

3.令 $C=C-a$， $U=U-POS_a(D)$， $B=\{a\}$， $\partial_B(x_i)=\partial_a(x_i)$

4.当 $C\neq\varnothing$ 时，对于 $\forall b\in C$：

4.1. $B=B\cup\{b\}$， $C=C-\{b\}$

4.2. 对 $\forall x_i\in U$， $\forall x_j\in\partial_B(x_i)$ 计算是否满足 $(x_i,x_j)\in TOL(B)$，不满足时 $\partial_B(x_i)=\partial_B(x_i)-\{x_j\}$；

4.3. 计算 $POS_B(D)$， $POS_C(D)=POS_C(D)\cup POS_B(D)$， $U=U-POS_B(D)$

5.输出正域集合 $POS_C(D)$。

3.2. 正域的快速计算

定理 2 [13]给定一个不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$，X_1 和 X_2 是 C 的两个属性子集，其中 $X_1\subseteq X_2\subseteq C$，可以得到 $POS_{X_1}(D)\subseteq POS_{X_2}(D)$。

定理 3 [13]给定一个不完备决策系统 $IDS=(U,AT=C\cup D,V,f)$，X_1 和 X_2 是 C 的两个属性子集，其中 $X_1\subseteq X_2\subseteq C$，对于 $\forall x_i\in U$，$x_i\in POS_{X_1}(D)$，那么 $x_i\in POS_{X_2}(D)$。

定理 2 和定理 3 表明在当任意一个对象属于较小属性子集的正域时，则必属于包含整个较小属性子集的较大属性子集的正域。

基于上述定理，表 2 给出不完备决策系统下快速正域计算算法(Fast Positive Domain Computing Algorithms inIncomplete Decision Systems, FPDCA-IDS)。

4. 实验分析

本实验选取八组 UCI 数据集进行求取正域效率对比，详细信息如表 3 所示。在验证算法有效性之前，使用 WEKA3.8 对数据集进行离散化预处理。本节进行的所有的实验都是在一台带有 MacOS13.4 Ventura、8 核、M1 芯片和 16GB 内存的笔记本电脑上进行。算法在 Pycharm2023 开发工具是使用 Python 编写，实验图使用 Python 绘制。

Table 3. Dataset information
表 3. 实验使用的数据集

序号	名称	对象数	特征数	分类数
1	Balance Scale	625	4	5
2	BreastCancer	699	9	2
3	Car Evaluation	1728	6	4
4	Dermatology	366	33	6
5	Iris	150	4	3
6	MONK’s Problems	432	6	2
7	Tic-Tac-Toe	958	9	2
8	Wine	178	13	3

4.1. 纬度效率验证

如图 1 所示为本文所提算法 FPDCA-IDS 与经典求取正域的算法 PDCA-IDS 在不完备决策系统下改变数据集的维度时(即改变属性个数的多少时)算法的执行时间曲线。红色球形折线为本文提出的算法 FPDCA-IDS，蓝色三角折线为传统的计算正域的方法 PDCA-IDS。横坐标为属性的个数，纵坐标为运行所需要的时间单位为秒。

从图 1 中可以看出，当数据集的属性个数都较少时算法的效率差别不大。当随着属性的个数逐渐增加时，两种算法的执行时间均有上升，但是 FPDCA-IDS 算法的折线始终保持在 PDCA-IDS 算法折线的下方，也就是 FPDCA-IDS 算法一直保持较好的效率。尤其是当属性的个数不断增加时，两种算法直接的差距更为明显，本文提出的算法优势更大。并且本文提出的算法随着属性个数的增加变化比较均匀，稳定性更好。因此，相对来说可以证明本文提出的算法更加适用于维度较高的数据集。

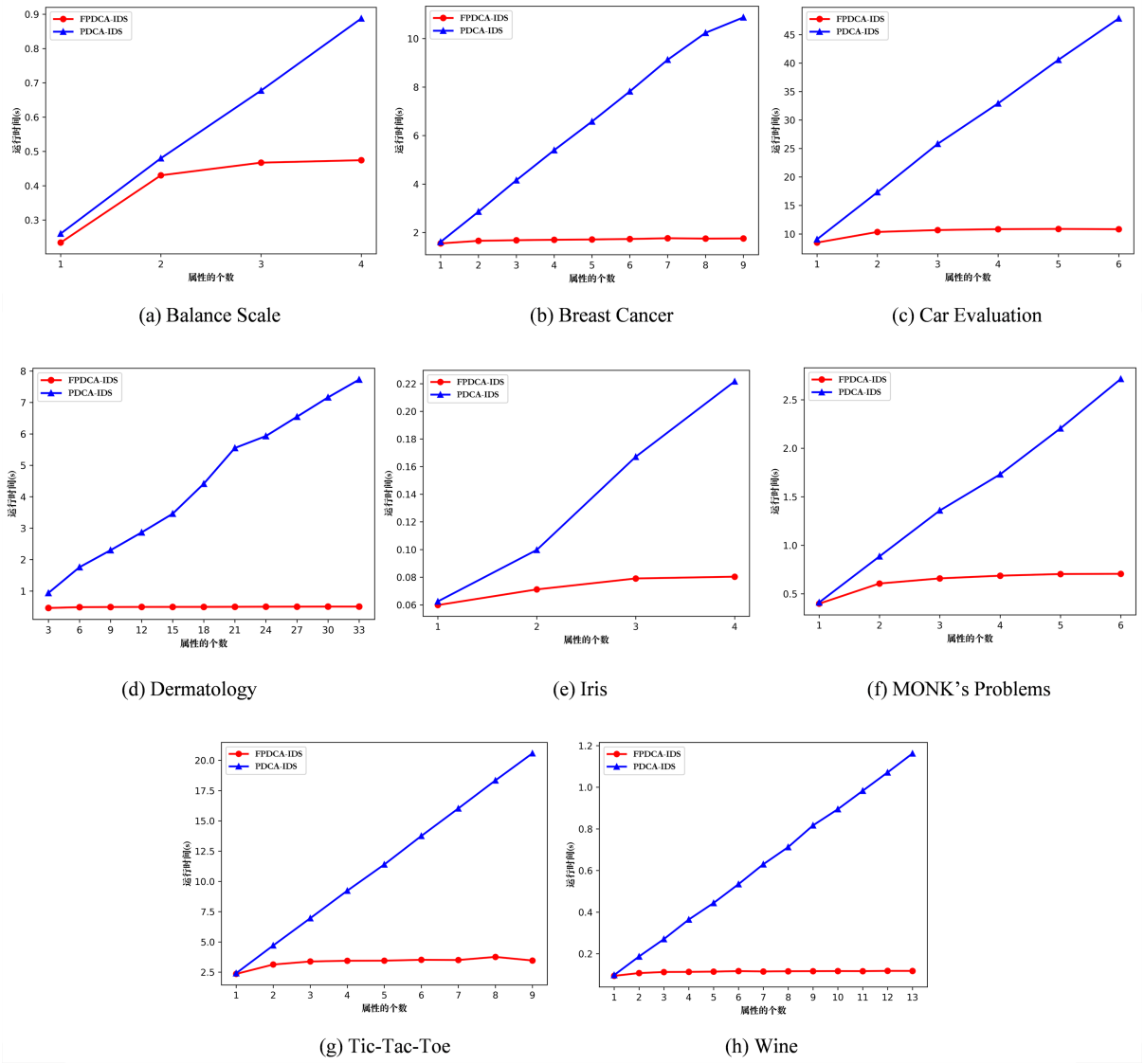

Figure 1. Comparison of efficiency in longitude
图 1. 维度下的效率对比

4.2. 规模效率验证

如图 2 所示为本文所提算法 FPDCA-IDS 与经典求取正域的算法 PDCA-IDS 在不完备决策系统下改变数据集的规模时(即改变样本对象个数的时)算法的执行时间曲线。红色球形折线为本文提出的算法 FPDCA-IDS，蓝色三角折线为传统的计算正域的方法 PDCA-IDS。横坐标为样本对象的比例，纵坐标为运行所需要的时间单位为秒。

从图 2 中可以看出，当数据集的样本对象个数较少时算法的效率差别不大。当随着属性的个数逐渐增加时，两种算法的执行时间均有上升，但是本文提出的算法随着样本对象比例的增加变化比较均匀，稳定性更好。本文提出的 FPDCA-IDS 算法的折线始终保持在 PDCA-IDS 算法折线的下方，也就是 FPDCA-IDS 算法一直保持较好的效率。当样本对象比例较高时，两种算法直接的差距更为的明显，本文提出的算法优势更大。因此，相对来说可以证明本文提出的算法更加适用于大规模的数据集。

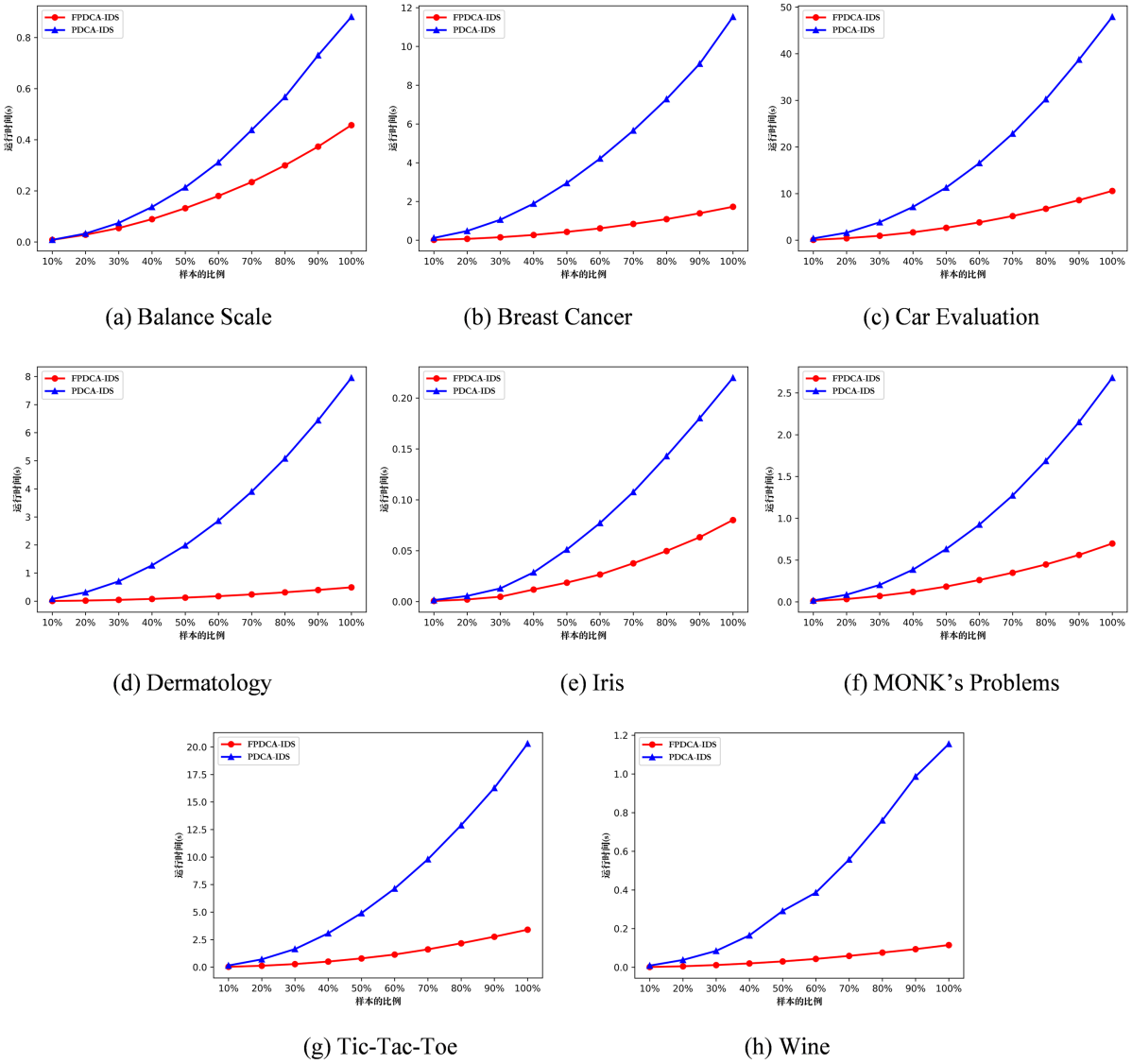

Figure 2. Comparison of efficiency in magnitude
图 2. 规模下的效率对比

5. 结论

在本文中，首先通过分析相容类的单调性以及正域在属性集合下的规律，进一步提出了一种时间复杂度较低的计算相容类和正域的算法。本文所提算法 FPDCA-IDS 与传统的计算正域的算法 PDCA-IDS 思想相比，减少了多次迭代次数和属性值的比较次数，能够有效的避免了时间复杂度过高的问题，可以大幅度提高算法的计算效率。本文最后通过选取的八组 UCI 数据集上进行算法有效性验证。从实验结果可以看出，本文所提算法相比于传统的计算正域的算法效率较高并且更加的稳定，尤其是更适用于大规模数据集以及维度较高的数据集。

基金项目

本文受烟台市科技计划项目(编号：2022XDRH016)的资助。

参考文献

[1] Pawlak, Z. (1982) Rough Sets. *International Journal of Computer and Information Sciences*, **11**, 341-356. https://doi.org/10.1007/BF01001956

[2] Liang, J.Y. and Xu, Z.B. (2002) The Algorithm on Knowledge Reduction in Incomplete Information Systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **10**, 95-103. https://doi.org/10.1142/S021848850200134X

[3] Meng, Z.Q. and Shi, Z.Z. (2009) A Fast Approach to Attribute Reduction in Incomplete Decision Systems with Tolerance Relation-Based Rough Sets. *Information Sciences*, **179**, 2774-2793. https://doi.org/10.1016/j.ins.2009.04.002

[4] Dai, J., Wei, B., Zhang, X., *et al*. (2017) Uncertainty Measurement for Incomplete Interval-Valued Information Systems Based on α-Weak Similarity. *Knowledge-Based Systems*, **136**, 159-171. https://doi.org/10.1016/j.knosys.2017.09.009

[5] Yang, X.B., Liang, S.C., Yu, H.L., *et al*. (2019) Pseudo-Label Neighborhood Rough Set: Measures and Attribute Reductions. *International Journal of Approximate Reasoning*, **105**, 112-129. https://doi.org/10.1016/j.ijar.2018.11.010

[6] Wang, C.Z., Huang, Y., Shao, M.W., *et al*. (2019) Fuzzy Rough Set-Based Attribute Reduction Using Distance Measures. *Knowledge-Based Systems*, **164**, 205-212. https://doi.org/10.1016/j.knosys.2018.10.038

[7] Tiwari, A.K., Shreevastava, S., Som, T., *et al*. (2018) Tolerance-Based Intuitionistic Fuzzy-Rough Set Approach for Attribute Reduction. *Expert Systems with Applications*, **101**, 205-212. https://doi.org/10.1016/j.eswa.2018.02.009

[8] Ali, A., Ali, M.I. and Rehman, N. (2019) Soft Dominance Based Rough Sets with Applications in Information Systems. *International Journal of Approximate Reasoning*, **113**, 171-195. https://doi.org/10.1016/j.ijar.2019.06.009

[9] Du, W.S. and Hu, B.Q. (2016) Dominance-Based Rough Set Approach to Incomplete Ordered Information Systems. *Information Sciences*, **346**, 106-129. https://doi.org/10.1016/j.ins.2016.01.098

[10] Kryszkiewicz, M. (1998) Rough Set Approach to Incomplete Information Systems. *Information Sciences*, **112**, 39-49. https://doi.org/10.1016/S0020-0255(98)10019-1

[11] Meng, Z.Q. and Shi, Z.Z. (2009) A Fast Approach to Attribute Reduction in Incomplete Decision Systems with Tolerance Relation-Based Rough Sets. *Information Sciences*, **179**, 2774-2793. https://doi.org/10.1016/j.ins.2009.04.002

[12] Peng, X., Wang, P., Xia, S., *et al*. (2022) FNC: A Fast Neighborhood Calculation Framework. *Knowledge-Based Systems*, **252**, 504-521. https://doi.org/10.1016/j.knosys.2022.109394

[13] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.