Hans Journal of Data Mining 数据挖掘, 2023, 13(4), 327-334
Published Online October 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.134032

Hans汉斯

序决策系统下基于图顶点最小覆盖的属性约简

战柏成

烟台大学计算机与控制工程学院，山东 烟台

收稿日期：2023年9月2日；录用日期：2023年10月2日；发布日期：2023年10月11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。属性约简是数据预处理的重要步骤之一，能够减少属性维度和计算复杂度，提高分类性能和可解释性。传统的属性约简方法主要基于信息论、统计学或启发式算法，存在不足之处。本文提出了一种基于图顶点最小覆盖的序决策系统属性约简方法，利用图来建模属性之间的依赖关系，使属性约简算法和图论知识相结合。实验结果表明，本文方法在多个数据集上具有较好的约简效果和分类性能，具有良好的可解释性和可视化效果。

关键词

粗糙集，序决策系统，图顶点最小覆盖理论，属性约简

Attribute Reduction Based on Minimum Cover of Graph Vertices in Ordered Decision System

Baicheng Zhan

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

Received: Sep. 2nd, 2023; accepted: Oct. 2nd, 2023; published: Oct. 11th, 2023

Abstract

In today's Internet era, the data dimension has grown more dramatically, and how to extract useful information from high-dimensional data has become a big problem. Attribute reduction is one of the important steps of data preprocessing, which can reduce the attribute dimension and computational complexity, and improve the classification performance and interpretability. Traditional attribute reduction methods are mainly based on information theory, statistics or heuristic

文章引用: 战柏成. 序决策系统下基于图顶点最小覆盖的属性约简[J]. 数据挖掘, 2023, 13(4): 327-334.
DOI: 10.12677/hjdm.2023.134032

algorithms, which are shortcomings. In this paper, we propose a method based on the minimum coverage of graph vertices to model the dependencies between properties and combining the attribute reduction algorithm and graph theory knowledge. Experimental results show that the present method has good reduction and classification performance on multiple datasets, with good interpretability and visualization.

Keywords

Rough Set, Order Decision System, Graph Vertex Minimum Covering Theory, Attribute Reduction

1. 引言

基于集合论，波兰学者 Pawlak 于 1982 年提出了粗糙集理论[1]，粗糙集理论是用于处理不精确、不一致、不完备信息和知识的有效工具，其主要思想是通过上下近似来表示数据中的确定性与不确定性，从而推导出相应的决策规则。属性约简[2] [3] [4] [5]作为粗糙集理论数据预处理中的一个重要步骤，目的是减少属性维度和计算复杂度，做到数据降维。目前，许多学者对属性约简作了深入的研究并取得了很多成果[6] [7] [8] [9] [10]，但目前的研究大多以在传统粗糙集理论不断改进启发式属性约简或差别矩阵算法为主[11] [12] [13]。

超图是图论中的一种扩展，能够表示节点之间的多重关系，被广泛应用于数据分析、图像分割、社交网络等领域。超图应用于数据分析主要包括超图聚类、超图分析、超图可视化等方面。超图理论将数据点作为节点，利用超图表达样本之间的关系。超图分析方法将数据点和特征作为节点，利用超图表达数据之间的依赖关系，实现数据分析，能够将复杂数据映射到超图上，展示数据之间的关系和特征，提高数据的可视化效果和理解性。

本文从超图理论的角度出发，将超图理论的最小顶点覆盖算法与序决策系统相结合，提高约简效率。经实验证明本文所提算法具备一定的有效性。

2. 基本概念

2.1. 序决策系统

设一个四元组 $CIS=(U,AT,V,f)$ 为一个信息系统，如果在信息系统 CIS 中的某一属性的值域上存在偏序关系，则称该属性为一个准则。当 CIS 中每个属性都为一个准则时，该系统为一个序决策系统 $OIS=(U,AT,V,f)$。其中：

- U：表示所有对象的集合，称为论域。
- AT：包含条件属性集合 C 和决策属性集合 D。
- V：表示条件属性集合 C 和决策属性集合 D 的值域。
- f：代表一个映射函数 $f:U\times AT\to V$，论域中的每一个对象在条件属性和决策属性上对应一个属性值，即 $f(x_i,c_k)=c_k(x_i)$。

定义 1 [14] 在序决策系统 $OIS=(U,AT,V,f)$ 中，对于 $\forall\mathbb{Q}\subseteq C$，有优势关系定义如下：

$$Dominance(\mathbb{Q})=\left\{(x_i,x_j)\in U\times U\middle|\forall q\in\mathbb{Q}, f(x_i,q)\geq f(x_j,q)\right\} \tag{1}$$

且满足 $Dominance(\mathbb{Q})=\cap_{c\in B}Dominance(\{c\})$。其中，$f(x_i,q)\geq f(x_j,q)$表示在属性 q 下，对象 x_i 优于对象 x_j。

优势关系是一个满足自反性、传递性和不对称性的偏序关系。因此，优势关系 $Dominance(\mathbb{Q})$ 可以导出论域在条件属性子集上的覆盖 $U/Dominance(\mathbb{Q})=U/\mathbb{Q}=\left\{[x_i]_{\mathbb{Q}}^{\geq}\middle|x_i\in U\right\}$。其中，$[x_i]_{\mathbb{Q}}^{2}$ 称之为优势类，且 $[x_i]_{\mathbb{Q}}^{\geq}=\left\{x_j\in U\middle|(x_j,x_i)\in Dominance(\mathbb{Q})\right\}$。在决策属性上，同样有一个覆盖表示为 $U/D=\left\{[x_i]_D^{\geq}\middle|x_i\in U\right\}=\left\{D_1^{\geq},D_2^{\geq},\cdots,D_n^{\geq}\right\}$，其中 $[x_i]_D^{\geq}=\left\{x_j\in U\middle|(x_j,x_i)\in Dominance(D)\right\}$ 称之为决策类。

定义 2 [14] 在序决策系统 $OIS=(U,AT,V,f)$ 中，对于 $\forall\mathbb{Q}\subseteq C$，有下、上近似定义如下：

$$\underline{Dominance}_{\mathbb{Q}}\left(D_i^{\geq}\right)=\left\{x_i\in U\middle|[x_i]_{\mathbb{Q}}^{\geq}\subseteq D_i^{\geq}\right\} \tag{2}$$

$$\overline{Dominance}_{\mathbb{Q}}\left(D_i^{\geq}\right)=\cup_{x_i\in D_i^{\geq}}[x_i]_{\mathbb{Q}}^{\geq} \tag{3}$$

且边界域表示为 $BND_{\mathbb{Q}}\left(D_i^{\geq}\right)=\overline{Dominance}_{\mathbb{Q}}\left(D_i^{\geq}\right)-\underline{Dominance}_{\mathbb{Q}}\left(D_i^{\geq}\right)$。

2.2. 超图

超图理论是对传统图论的扩展和推广。在传统图论中，图由节点和边组成，而超图引入了超边的概念，以允许节点之间的多重关系。在超图中，超边可以连接多个节点，而传统图中的边只能连接两个节点。超边可以连接任意数量的节点，从而更好地捕捉节点之间的复杂关系。每个节点和超边可以具有任意类型的属性，这使得超图可以应用于各种实际问题。

定义 3 [15] 设超图为一个二元组 $H=(V,E)$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 为顶点集合，$E=\{e_1,e_2,\cdots,e_m\}$ 为超边集合。

如图 1 所示为一个超图，共包含 13 个顶点和 8 条超边。

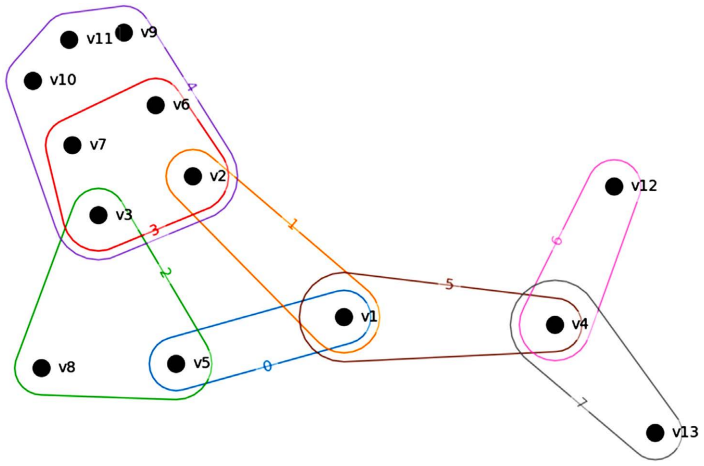


Figure 1. The hypergraph example
图 1. 超图示例

3. 基于序决策系统的下近似属性约简算法

序决策系统的下近似约简定义如下：

定义 4 [14] 在序决策系统 $OIS=(U,AT,V,f)$ 中，对于 $\forall\mathbb{Q}\subseteq C$，$U/D=\left\{D_1^{\geq},D_2^{\geq},\cdots,D_n^{\geq}\right\}$。当 $RED\subseteq C$ 为序决策系统的下近似约简，当且仅当满足以下条件：

1) 对于 $\forall D_i^{\geq} \in U/D$，有 $\underline{Dominance}_{RED}\left(D_i^{\geq}\right) = \underline{Dominance}_{C}\left(D_i^{\geq}\right)$；

2) 对于 $\forall red \in RED$，$\exists D_i^{\geq} \in U/D$，有 $\underline{Dominance}_{red}\left(D_i^{\geq}\right) \neq \underline{Dominance}_{RED}\left(D_i^{\geq}\right)$。

条件(1)为联合充分性，条件(2)为个体必要性。

定义 5 [14] 在序决策系统 $OIS=(U,AT,V,f)$ 中，对于 $\forall \mathbb{Q} \subseteq C$，$U/D=\left\{D_1^{\geq},D_2^{\geq},\cdots,D_n^{\geq}\right\}$。$Matrix_{OIS}=\left(M_{OIS}\left(x_i,x_j\right)\right)$ 为一个规模为 $|U|^2$ 的差别矩阵。且满足以下条件：

$$M_{OIS}\left(x_i,x_j\right)=\begin{cases}\left\{c_k : c_k \in C\right\}, & \left(x_i,x_j\right) \in Condition \\ \varnothing, & \text{otherwise}\end{cases} \tag{4}$$

其中，$Condition=\cup_{D_i^{\geq} \in U/D}\left\{\left(x_i,x_j\right) \middle| x_i \notin Dominance_C\left(D_i^{\geq}\right), x_j \in Dominance_C\left(D_i^{\geq}\right)\right\}$。

定义 6 [14] 在序决策系统 $OIS=(U,AT,V,f)$ 中，对于 $\forall \mathbb{Q} \subseteq C$，$U/D=\left\{D_1^{\geq},D_2^{\geq},\cdots,D_n^{\geq}\right\}$。$Matrix_{OIS}=\left(M_{OIS}\left(x_i,x_j\right)\right)$ 为一个规模为 $|U|^2$ 的差别矩阵，其差别函数定义如下：

$$F_{OIS}\left(x_i,x_j\right)=\wedge\left(\vee M_{OIS}\left(x_i,x_j\right)\right) \tag{5}$$

基于上述定义，给出序决策系统下经典差别矩阵算法(Classical discernibility matrix algorithm in ordered decision system, CDM-ODS) (表 1)。

Table 1. Classical discernibility matrix algorithm in ordered decision system (CDM-ODS)
表 1. 序决策系统下经典差别矩阵算法

输入： 序决策系统 $OIS=(U,AT,V,f)$。

输出： 所有约简结果集合 RED。

1) 初始化：$RED=\varnothing$；

2) 根据条件属性集合 C 和决策属性 D 计算序决策系统下 $OIS=(U,AT,V,f)$ 每个对象的优势类 $[x_i]_C^{\geq}$ 和决策类 $U/D=\left\{D_1^{\geq},D_2^{\geq},\cdots,D_n^{\geq}\right\}$；

3) 对决策类 $U/D=\left\{D_1^{\geq},D_2^{\geq},\cdots,D_n^{\geq}\right\}$ 中的每一个 $D_i^{\geq}$ 计算下近似集合 $\underline{Dominance}_{C}\left(D_i^{\geq}\right)$；

4) 根据下近似 $\underline{Dominance}_{C}\left(D_i^{\geq}\right)$ 计算差别矩阵 $Matrix_{OIS}$；

5) 计算差别函数 $F_{OIS}\left(x_i,x_j\right)$；

6) 将差别函数 $F_{OIS}\left(x_i,x_j\right)$ 转化为极小析取范式；

7) 输出所有约简集合 RED。

4. 基于图顶点最小覆盖的属性约简算法

差别矩阵算法进行属性约简的过程与图论中寻找最小顶点覆盖集之间存在共同点。因此针对这种情况，可以将序决策系统构造出的差别矩阵生成一张超图，在超图上进行求解顶点最小覆盖。

定义 7 设二元组 $H=(V,E)$ 为一个超图，V 顶点集合，E 是超边集合。在超图 $H=(V,E)$ 中定义如下布尔函数：

$$F_H=\wedge\left\{\vee e : e \in E\right\} \tag{6}$$

对该函数求解极小析取范式所得结果就是最小顶点覆盖集合。

定义 8 在序决策系统 $OIS=(U,AT,V,f)$ 中， $Matrix_{OIS}=\left(M_{OIS}\left(x_i,x_j\right)\right)$ 为序决策系统生成的差别矩阵，由该系统差别矩阵导出的超图定义如下：

$$H_{OIS}=(V,E) \tag{7}$$

其中，顶点集合 $V=C$，超边集合 $E=M_{OIS}\left(x_i,x_j\right)$。

定义 9 在序决策系统 $OIS=(U,AT,V,f)$ 中， $Matrix_{OIS}=\left(M_{OIS}\left(x_i,x_j\right)\right)$ 为序决策系统生成的差别矩阵，由该系统差别矩阵导出的超图为 $H_{OIS}=(V,E)$。顶点集合 $V=C$，则 V 中顶点的度表示为 $Degree_H\left(v_i\right)$，顶点的度表示为与顶点相关联的超边数。

定理 1 在序决策系统 $OIS=(U,AT,V,f)$ 中， $H_{OIS}=(V,E)$ 为序决策系统的差别矩阵生成的超图，则有 $RED_{OIS}=Vertex\left(H_{OIS}\right)$。

其中， $Vertex\left(H_{OIS}\right)$ 代表超图 $H_{OIS}=(V,E)$ 的定点最小覆盖。

证明：

要证明 $RED_{OIS}=Vertex\left(H_{OIS}\right)$ 首先假设 $Matrix_{OIS}=\left(M_{OIS}\left(x_i,x_j\right)\right)$ 为该系统生成的差别矩阵。通过定义 8 可知诱导超图 $H_{OIS}=(V,E)$ 的超边集合 $E=M_{OIS}\left(x_i,x_j\right)\in Matrix_{OIS}$，因此有 $E\subseteq Matrix_{OIS}$。这将分为以下两种情况：

情况一： $E=Matrix_{OIS}$；

情况二： $E\subset Matrix_{OIS}$。

在情况一中，通过定义 7 可知显然 $F_H=\wedge\{\vee e:e\in E\}=F_{OIS}\left(x_i,x_j\right)=\wedge\left(\vee M_{OIS}\left(x_i,x_j\right)\right)$；在情况二中，由 $E\subseteq Matrix_{OIS}$ 可知 $Matrix_{OIS}=E\cup\{C\}$，即 $M_{OIS}\left(x_i,x_j\right)\in Matrix_{OIS}\subset C$，根据吸收律可得 $\left(\vee M_{OIS}\left(x_i,x_j\right)\right)\wedge(\vee C)=\vee M_{OIS}\left(x_i,x_j\right)$。综上得 $F_H=\wedge\{\vee e:e\in E\}=F_{OIS}\left(x_i,x_j\right)=\wedge\left(\vee M_{OIS}\left(x_i,x_j\right)\right)$。

基于上述定理，给出序决策系统下基于图顶点最小覆盖的属性约简算法(Attribute reduction algorithm based on graph vertex minimum cover in ordered decision system, ARGVMC-ODS) (表 2)。

Table 2. Attribute reduction algorithm based on graph vertex minimum cover in ordered decision system, ARGVMC-ODS
表 2. 序决策系统下基于图顶点最小覆盖的属性约简算法

输入： 序决策系统 $OIS=(U,AT,V,f)$。

输出： 所有约简结果集合 RED。

1) 初始化： $RED=\varnothing$；

2) 求序决策系统 $OIS=(U,AT,V,f)$ 的差别矩阵 $Matrix_{OIS}$；

3) 将差别矩阵 $Matrix_{OIS}$ 转化为 $H_{OIS}=(V,E)$；

4) 针对超图 $H_{OIS}=(V,E)$，若 $E\neq\varnothing$，重复 5 至 7：

5) 计算顶点的度 $Degree_H\left(v_i\right)$，按从小到大排序；

6) 选择 $v_0=\max\left\{Degree_H\left(v_i\right):v_i\in V-RED\right\}$，并使 $RED=RED\cup\{v_0\}$；

7) 计算 $E'=E-\bigcup_{v_i\in RED}d_{b_{MIDS}}\left(v_i\right)$，并使 $E=E'$；

8) 输出所有约简集合 RED。

5. 实验分析

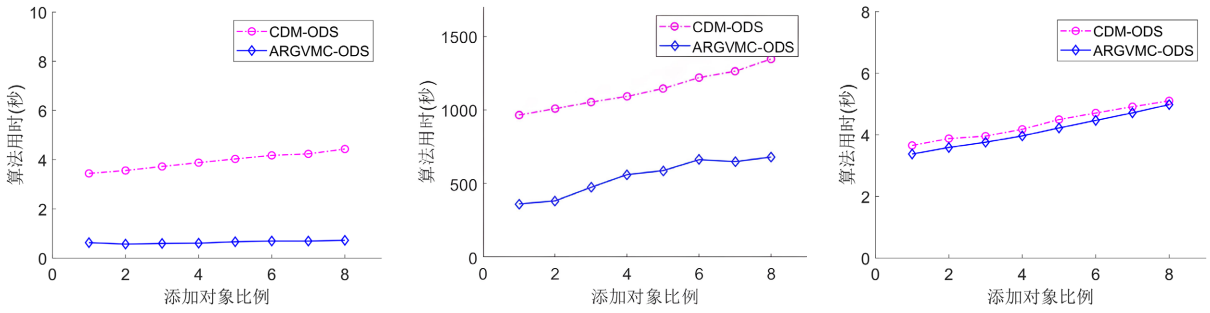
本实验选取八组 UCI 数据集进行约简效率和分类精度的对比，详细信息如表 3 所示。在验证算法有效性之前，使用 WEKA3.8 对数据集进行离散化预处理。本节进行的所有的实验都是在一台带有 MacOS13.0 Ventura、Intel(R) Core(TM) i9-9880H 八核 2.30 GHz CPU 和 16GB 内存的笔记本电脑上进行。算法在 Pycharm2023 开发工具是使用 Python 编写，实验图使用 MatlabR2021a 绘制。

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

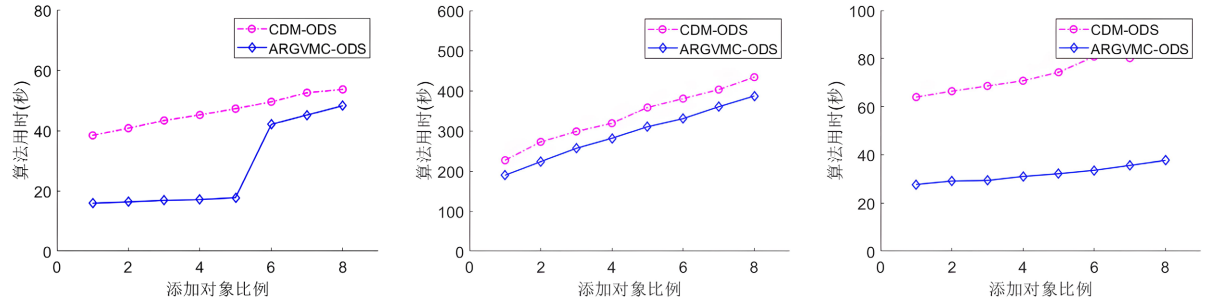
序号	名称	对象数	特征数	分类数
1	Wine	178	13	2
2	Abalone	4177	8	29
3	Car	1728	6	4
4	Breast Cancer	699	9	2
5	Concrete	1030	9	15
6	Connection	990	10	11
7	Yeast	1484	8	10
8	Zoo	101	16	2/3/7

5.1. 约简效率

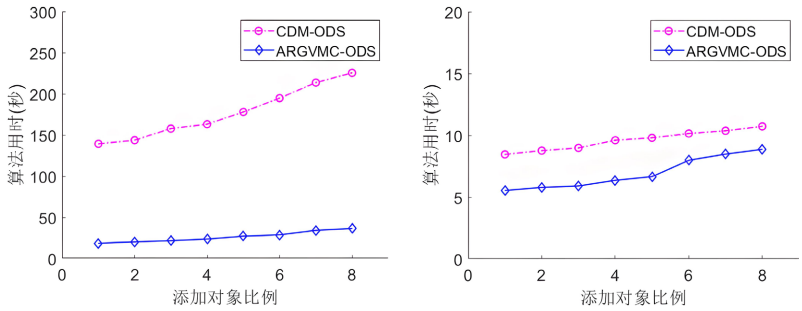
如图 2 所示为本文所提算法 ARGVMC-ODS 与经典差别矩阵算法 CDM-ODS 在序决策系统下按论域规模比例均匀增加时的算法执行时间曲线。从中可以看出，当论域中的对象数量比例较小时，算法效率差别不大。当对象的规模逐渐增大时，两种算法的执行时间均有上升，并且经典差别矩阵算法 CDM-ODS



(a) Wine (b) Abalone (c) Car



(d) Breast Cancer (e) Concrete (f) Connectionist Bench



(g) Yeast (h) Zoo

Figure 2. Comparison of reduction efficiency
图 2. 约简效率对比

的执行时间始终最长。本文所提算法 ARGVMC-ODS 与之相比，整体运行时间都有所缩短，且变化较均匀。因此，本文所提算法 ARGVMC-ODS 相对于经典差别矩阵算法 CDM-ODS 的约简效率较高，在大规模数据集上优势较为明显。

5.2. 分类精度

如图 3 与图 4 所示为本文所提算法 ARGVMC-ODS 与经典差别矩阵算法 CDM-ODS 在序决策系统下在不同数据集上分别在 KNN 与 SVM 分类器上的分类精度对比。在本实验中采取十倍交叉验证的方法计算分类精度。其中，经典差别矩阵算法 CDM-ODS 能够得出所有约简结果，而本文所提算法 ARGVMC-ODS 能够得出最短结果，因此对经典差别矩阵算法 CDM-ODS 的约简结果取平均值。从图中可以看出，本文所提算法在实验数据集上有较高的分类精度。

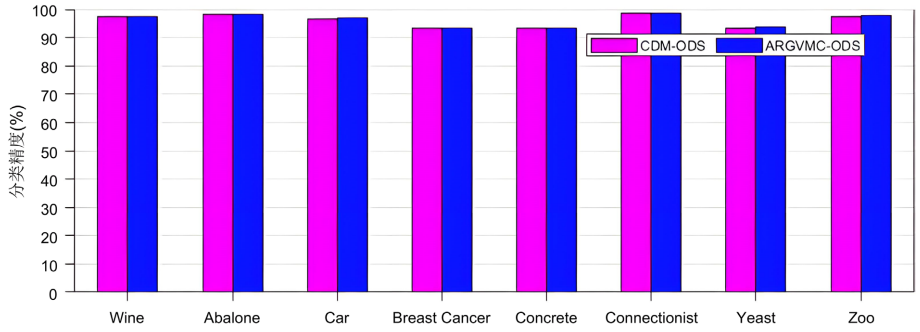


Figure 3. Comparison of classification accuracy (KNN)
图 3. 分类精度对比(KNN)

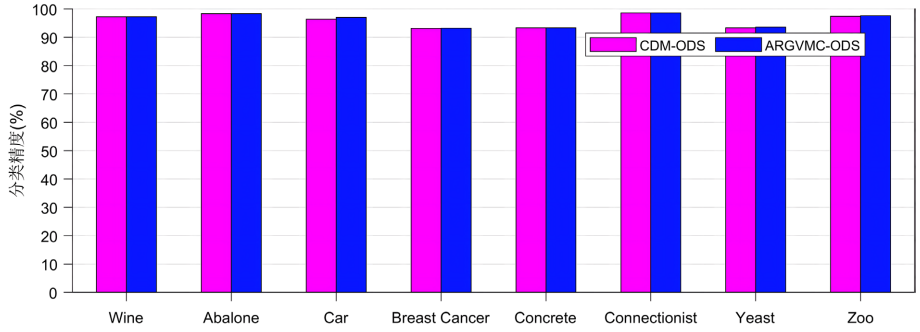


Figure 4. Comparison of classification accuracy (SVM)
图 4. 分类精度对比(SVM)

6. 结论

在本文中，首先通过寻找最小顶点覆盖和差别矩阵属性约简的共性，进一步提出在序决策系统回顾下基于顶点最小覆盖的属性约简算法，将图论与粗糙集理论相融合。本文所提算法 ARGVMC-ODS 与差别矩阵算法思想相比，无需计算最小析取范式，避免了时间复杂度过高的问题，能够提高算法效率。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差别矩阵算法效率较高。

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] Chen, Q., Huang, M., Wang, H. and Xu, G. (2022) A Feature Discretization Method Based on Fuzzy Rough Sets for High-Resolution Remote Sensing Big Data under Linear Spectral Model. *IEEE Transactions on Fuzzy Systems*, **30**, 1328-1342. https://doi.org/10.1109/TFUZZ.2021.3058020

[3] Dai, J.H., Hu, H., Zheng, G.J., *et al*. (2016) Attribute Reduction in Interval-Valued Information Systems Based on Information Entropies. *Frontiers of Information Technology & Electronic Engineering*, **17**, 919-928. https://doi.org/10.1631/FITEE.1500447

[4] Sun, L., Yin, T., Ding, W., Qian, Y. and Xu, J. (2022) Feature Selection with Missing Labels Using Multilabel Fuzzy Neighborhood Rough Sets and Maximum Relevance Minimum Redundancy. *IEEE Transactions on Fuzzy Systems*, **30**, 1197-1211. https://doi.org/10.1109/TFUZZ.2021.3053844

[5] Bao, H., Wu, W.Z., Zheng, J.W. and Li, T.J. (2021) Entropy Based Optimal Scale Combination Selection for Generalized Multi-Scale Information Tables. *International Journal of Machine Learning and Cybernetics*, **12**, 1427-1437. https://doi.org/10.1007/s13042-020-01243-y

[6] Yang, X.B., Qi, Y., Yu, D.J., Yu, H.L. and Yang, J.Y. (2015) α-Dominance Relation and Rough Sets in Interval-Valued Information Systems. *Information Sciences*, **294**, 334-347. https://doi.org/10.1016/j.ins.2014.10.003

[7] Qian, Y.H., Liang, X.Y., Wang, Q., Liang, J.Y., Liu, B., Skowron, A., Yao, Y.Y., Ma, J.M. and Dang, C.Y. (2018) Local Rough Set: A Solution to Rough Data Analysis in Big Data. *International Journal of Approximate Reasoning*, **97**, 38-63. https://doi.org/10.1016/j.ijar.2018.01.008

[8] Shu, W.H. and Qian, W.B. (2015) An Incremental Approach to Attribute Reduction from Dynamic Incomplete Decision Systems in Rough Set Theory. *Data & Knowledge Engineering*, **100**, 116-132. https://doi.org/10.1016/j.datak.2015.06.009

[9] Yao, Y.Y. and Zhang, X.Y. (2017) Class-Specific Attribute Reducts in Rough Set Theory. *Information Sciences*, **418-419**, 601-618. https://doi.org/10.1016/j.ins.2017.08.038

[10] 张楠, 苗夺谦, 岳晓冬. 区间值信息系统的知识约简[J]. 计算机研究与发展, 2010, 47(8): 1362-1371.

[11] Skowron, A. and Rauszer, C. (1992) The Discernibility Matrices and Functions in Information Systems. In: Słowiński, R., Ed., *Intelligent Decision Support*, Springer, Dordrecht, 331-362. https://doi.org/10.1007/978-94-015-7975-9_21

[12] Huang, Z.H., Li, J.J., Dai, W.Z. and Lin, R. (2019) Generalized Multi-Scale Decision Tables with Multi-Scale Decision Attributes. *International Journal of Approximate Reasoning*, **115**, 194-208. https://doi.org/10.1016/j.ijar.2019.09.010

[13] Huang, Y., Li, T., Luo, C., *et al*. (2017) Matrix-Based Dynamic Updating Rough Fuzzy Approximations for Data Mining. *Knowledge-Based Systems*, **119**, 273-283. https://doi.org/10.1016/j.knosys.2016.12.015

[14] 孙祖文, 唐玉凯. 序决策系统下近似约简的启发式算法[J]. 计算机科学与应用, 2021, 11(1): 113-120.

[15] Bretto, A. (2013) Hypergraph Theory. Springer, Cham.