Hans Journal of Data Mining 数据挖掘, 2023, 13(3), 213-221
Published Online July 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.133021

Hans汉斯

广义多粒度粗糙集特征选择算法研究

梁晓敏

烟台大学计算机与控制工程学院，山东 烟台

收稿日期：2023年5月27日；录用日期：2023年6月27日；发布日期：2023年7月5日

摘　要

随着信息技术的迅猛发展，产生了大量的数据，这些数据体量巨大、形式多样、产生迅速、价值密度低、商业价值高。如何使这些数据对人类社会的进步产生积极影响是一个难题。粗糙集理论可以直接对数据进行降维处理，发现数据中的隐含知识，促进社会进步。经典粗糙集理论基于单个二元关系，缺乏灵活性和普遍性，基于多个二元关系的粗糙集理论可以解决上述难题，因此，本文主要针对广义多粒度粗糙集进行了研究，引入元启发式算法，提出通过元启发式算法(蚁群算法)实现广义多粒度粗糙集特征选择算法。通过实验结果看出本文所提算法可以对数据集起到降维效果且得到的特征子集的分类精度和原数据集基本保持一致。

关键词

粒计算，特征选择，广义多粒度粗糙集，二元关系

Researches on Feature Selection Algorithm for Generalized Multi-Granularity Rough Sets

Xiaomin Liang

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

Received: May 27th, 2023; accepted: Jun. 27th, 2023; published: Jul. 5th, 2023

Abstract

With the rapid development of information technology, a large amount of data has been generated, which is huge in volume, diverse in form, rapid in generation, low in value density, and high in commercial value. How to make these data have a positive impact on the progress of human society

文章引用: 梁晓敏. 广义多粒度粗糙集特征选择算法研究[J]. 数据挖掘, 2023, 13(3): 213-221.
DOI: 10.12677/hjdm.2023.133021

is a challenge. Rough set theory can directly reduce the dimensionality of the data, discover the implicit knowledge in the data, and promote the social progress. The classical rough set theory is based on a single binary relationship, which lacks flexibility and universality. The rough set theory based on multiple binary relationships can solve the above problems. Therefore, this paper mainly focuses on the generalized multi-granularity rough set and introduces the meta-heuristic algorithm, and proposes to implement the generalized multi-granularity rough set feature selection algorithm by the meta-heuristic algorithm (ant colony algorithm). The experimental results show that the proposed algorithm can reduce the dimensionality of the data set and the classification accuracy of the obtained feature subsets is basically consistent with the original data set.

Keywords

Granular Computing, Feature Selection, Generalized Multi-Granularity Rough Sets, Binary Relationships

1. 引言

粗糙集理论[1]是一个用于处理不确定的数学工具，经典粗糙集理论建立在单个二元关系上，不能用于处理具有多个二元关系的决策系统。Qian 等[2]提出了基于多个二元关系的决策系统，称为多粒度粗糙集理论。多粒度粗糙集模型分为乐观多粒度粗糙集模型和悲观多粒度粗糙集模型，由于乐观多粒度粗糙集模型的构造过于放松，悲观多粒度粗糙集模型的构造过于严苛，因此，Xu 等[3]提出了广义多粒度粗糙集模型。一些学者针对广义多粒度粗糙集理论进行了深入研究。Qian 等[4]构造了一个广义层次决策表，并将多粒度和序贯三支决策相结合，提出了广义层次多粒度序贯三支决策模型。Xu 等[5]通过考虑类与概念之间的相对和绝对定量信息，提出了两种广义多粒双定量决策理论粗糙集模型。Xu 等[6]针对局部广义多粒度邻域粗糙集模型，提出了动态更新近似的方法。张先韬等[7]给出了广义多粒度粗糙集约简的一些基本性质，给出 matlab 计算的过程及计算实例。

在已有研究中，广义多粒度粗糙集特征选择的研究并不完善，未有人通过元启发式算法进行广义多粒度粗糙集特征选择算法的研究。元启发式算法是启发式算法的改进，由于其有较好的泛化性、较强的通用性，现已被广泛应用于各个领域。因此，本文首先介绍了广义多粒度粗糙集的相关概念，然后详细介绍了蚁群算法的基础知识，在此基础上提出通过元启发算法(蚁群算法)实现对广义多粒度粗糙集特征选择算法的研究。实验结果表明，本文所提的算法可以对高维数据进行降维，并且得到的特征子集并没有降低原数据集的分类精度。

2. 基本概念

四元组 $DS=\left(U,AT=C\cup D,V,f\right)$ 为决策系统，其中 U 为论域，C 为条件属性集，D 为决策属性集，V 为 $a_k\in AT$ 的值域集，f 为映射函数。

定义 1 [8]给定决策系统 $DS=\left(U,AT=C\cup D,V,f\right)$，对 $X\subseteq U$，粒度集 $P=\left\{P_1,P_2,\cdots,P_m\right\}$，$P_i\subseteq C\left(1\le i\le m\right)$，通过特征函数 $S_X^{P_i}\left(x\right)$ 描述 X 和等价类 $\left[x\right]_{P_i}$ 之间的包含关系，特征函数 $S_X^{P_i}\left(x\right)$ 定义为：

$$S_X^{P_i}(x)=\begin{cases}1, & [x]_{P_i}\subseteq X\\ 0, & \text{其他}\end{cases}\quad(1\le i\le m),$$

其中 $S_X^{P_i}(x)$ 为 x 的特征函数。

定义 2 [8]给定决策系统 $DS=(U,AT=C\cup D,V,f)$，对 $X\subseteq U$，粒度集 $P=\{P_1,P_2,\cdots,P_m\}$，$P_i\subseteq C(1\le i\le m)$，$S_X^{P_i}(x)$ 是 x 的支持特征函数，$\beta\in(0.5,1]$，X 在粒度集 P 上的广义多粒度粗糙集上近似集和下近似集定义为：

$$\overline{\sum_{i=1}^{m}P_i}(X)_\beta=\left\{x\in U\mid\frac{\sum_{i=1}^{m}\left(1-S_{\sim X}^{P_i}(x)\right)}{m}>1-\beta\right\},$$

$$\underline{\sum_{i=1}^{m}P_i}(X)_\beta=\left\{x\in U\mid\frac{\sum_{i=1}^{m}S_X^{P_i}(x)}{m}\ge\beta\right\}.$$

定义 3 [8]给定决策系统 $DS=(U,AT=C\cup D,V,f)$，对 $X\subseteq U$，粒度集 $P=\{P_1,P_2,\cdots,P_m\}$，$P_i\subseteq C(1\le i\le m)$，$X$ 在粒度集 P 上的广义多粒度粗糙集正域、边界域以及负域定义为：

$$POS_{\underline{\sum_{i=1}^{m}P_i}}(X)_\beta=\underline{\sum_{i=1}^{m}P_i}(X)_\beta,$$

$$BND_{\underline{\sum_{i=1}^{m}P_i}}(X)_\beta=\overline{\sum_{i=1}^{m}P_i}(X)_\beta-\underline{\sum_{i=1}^{m}P_i}(X)_\beta,$$

$$NEG_{\underline{\sum_{i=1}^{m}P_i}}(X)_\beta=U-\left(POS_{\underline{\sum_{i=1}^{m}P_i}}(X)_\beta\cup BND_{\underline{\sum_{i=1}^{m}P_i}}(X)_\beta\right).$$

定义 4 [8]给定决策系统 $DS=(U,AT=C\cup D,V,f)$，不可分辨关系 $IND(D)$ 在 U 上导出的划分为 $U/D=\{D_1,D_2,\cdots,D_r\}(1\le r\le|U|)$，粒度集 $P=\{P_1,P_2,\cdots,P_m\}$，$P_i\subseteq C(1\le i\le m)$，$\beta\in(0.5,1]$，决策类集合 U/D 在粒度集 P 上的广义多粒度粗糙集上近似集和下近似集定义为：

$$\overline{\sum_{i=1}^{m}P_i}(U/D)_\beta=\left\{\overline{\sum_{i=1}^{m}P_i}(D_1)_\beta,\overline{\sum_{i=1}^{m}P_i}(D_2)_\beta,\cdots,\overline{\sum_{i=1}^{m}P_i}(D_r)_\beta\right\},$$

$$\underline{\sum_{i=1}^{m}P_i}(U/D)_\beta=\left\{\underline{\sum_{i=1}^{m}P_i}(D_1)_\beta,\underline{\sum_{i=1}^{m}P_i}(D_2)_\beta,\cdots,\underline{\sum_{i=1}^{m}P_i}(D_r)_\beta\right\}.$$

决策类集合 U/D 在粒度集 P 上的正域和边界域定义为：

$$POS_{\underline{\sum_{i=1}^{m}P_i}}(U/D)_\beta=\bigcup_{D_l\in U/D}\underline{\sum_{i=1}^{m}P_i}(D_l)_\beta,$$

$$BND_{\underline{\sum_{i=1}^{m}P_i}}(U/D)_\beta=\bigcup_{D_l\in U/D}\overline{\sum_{i=1}^{m}P_i}(D_l)_\beta-\bigcup_{D_l\in U/D}\underline{\sum_{i=1}^{m}P_i}(D_l)_\beta.$$

性质 1 给定决策系统 $DS=(U,AT=C\cup D,V,F,f)$，对 $\forall A\subseteq B\subseteq C$，粒度集 $P\_B=\{B_1,B_2,\cdots,B_m\}$，

$B_i \subseteq B(1 \le i \le m)$，$\bigcup_{i=1}^{m} B_i = B$，粒度集 $P\_A = \{A_1, A_2, \cdots, A_k\}$，$A_l \subseteq A(1 \le l \le k)$，$\bigcup_{l=1}^{k} A_l = A$，$\beta \in (0.5,1]$，可得 $POS_{\sum_{l=1}^{k} A_l}(U/D)_\beta \subseteq POS_{\sum_{i=1}^{m} B_i}(U/D)_\beta$ 或 $POS_{\sum_{l=1}^{k} A_l}(U/D)_\beta \supseteq POS_{\sum_{i=1}^{m} B_i}(U/D)_\beta$ 均不恒成立。

3. 广义多粒度粗糙集特征选择算法

元启发式算法包括遗传算法[9]、蜂群算法[10]、蚁群算法[11]等。接下来将详细介绍蚁群算法。

现实生活中，蚂蚁在觅食的过程中会在其经过的路径上留下信息素，后面的蚂蚁能感知到路径上的信息素，依据信息素指导自己的行为，选择具有信息素含量较多的路径可能性最大，也会留下信息素并对走过路径上的信息素加强。这样，大量蚂蚁组成的集体觅食行为就表现为对信息素正反馈的现象，进而逼近了最优路径。受现实生活中蚂蚁觅食的影响，Dorigo 等[11]提出了蚁群优化算法。Jensen 等[12]将蚁群优化算法用于粗糙集中的特征选择。Chen 等[13]将粗糙集中求取核属性集的方法融合到利用蚁群优化算法进行特征选择的算法中。特征选择的过程中，将单个条件属性看做一个节点，节点和节点之间的路径就是特征选择的过程，首先计算决策系统的核属性集，然后定义最大迭代次数，在每次迭代过程中给定一个由蚁群组成的搜索空间，蚁群中的每只人工蚂蚁从核属性集开始构造解，随机选择一个节点，再依据概率公式进行下一个节点的选择直到满足构造解的停止条件。每轮迭代结束后，进行信息素的更新，迭代过程结束后得到最优特征子集。下面将详细说明通过蚁群算法对决策系统进行特征选择的过程。

启发式信息

定义 5 [8]给定决策系统 $DS = (U, AT = C \cup D, V, f)$，粒度集 $P = \{P_1, P_2, \cdots, P_m\}$，$P_i \subseteq C(1 \le i \le m)$，对 $\beta \in (0.5,1]$，D 关于粒度集 P 在广义多粒度粗糙集下的依赖度定义为：

$$r(\sum_{i=1}^{m} P_i, D)_\beta = \frac{|POS_{\sum_{i=1}^{m} P_i}(U/D)_\beta|}{|U|}.$$

定义 6 给定决策系统 $DS = (U, AT = C \cup D, V, f)$，粒度集 $P\_C = \{C_1, C_2, \cdots, C_n\}$，$C_i \subseteq C(1 \le i \le n)$，$\bigcup_{i=1}^{n} C_i = C$，粒度集 $P\_B = \{B_1, B_2, \cdots, B_m\}$，$B_j \subseteq B(1 \le j \le m)$，$\bigcup_{j=1}^{m} B_j = B$，粒度集 $P\_A = \{A_1, A_2, \cdots, A_k\}$，$A_l \subseteq A(1 \le l \le k)$，$\bigcup_{l=1}^{k} A_l = A$，$A \subset B \subseteq C$，$\beta \in (0.5,1]$，若属性集 B 为 DS 的广义多粒度属性约简，那么 B 应该满足如下条件：

1) $r\left(\sum_{j=1}^{m} B_j, D\right)_\beta = r\left(\sum_{i=1}^{n} C_i, D\right)_\beta$；

2) $\forall A \subset B$，$r\left(\sum_{l=1}^{k} A_l, D\right)_\beta \ne r\left(\sum_{j=1}^{m} B_j, D\right)_\beta$。

定义 7 给定决策系统 $DS = (U, AT = C \cup D, V, f)$，粒度集 $P\_B = \{B_1, B_2, \cdots, B_m\}$，$B_i \subseteq B(1 \le i \le m)$，$\bigcup_{i=1}^{m} B_i = B$，粒度集 $P\_DB = \{B_{d1}, B_{d2}, \cdots, B_{dm}\}$，$B_l \subseteq B - \{b\}$ $(d1 \le l \le dm)$，$\bigcup_{l=d1}^{dm} B_l = B - \{b\}$，$B - \{b\} \subseteq B \subseteq C$，$\beta \in (0.5,1]$，对 $\forall b \in B$ 的内部属性重要度定义为：

$$Sig_{inner}(b, B, D)_\beta = r\left(\sum_{i=1}^{m} B_i, D\right)_\beta - r\left(\sum_{l=d1}^{dm} B_l, D\right)_\beta.$$

当 $Sig_{inner}(b, B, D)_\beta = 0$，说明属性 b 是不重要的，当 $Sig_{inner}(b, B, D)_\beta \ne 0$，说明属性 b 是不可缺少的。

因此可以将核属性集定义为：

$$core(C)=\left\{b\in C \mid Sig_{inner}(b,C,D)_{\beta}\neq 0\right\}.$$

性质 2 广义多粒度粗糙集理论中，满足核属性集是约简集的交集，即：

$$core(C)=\cap RED(C).$$

证明：

给定决策系统 $DS=(U,AT=C\cup D,V,f)$，粒度集 $P\_C=\{C_1,C_2,\cdots,C_n\}$，$C_i\subseteq C(1\le i\le n)$，$\bigcup_{i=1}^{n}C_i=C$，$\beta\in(0.5,1]$。

1) 设 $a\notin\cap RED(C)$，$\exists R\in RED(C)$ 使得 $a\notin R$，满足 $r\left(\sum_{j=1}^{m}R_j,D\right)_{\beta}=r\left(\sum_{i=1}^{n}C_i,D\right)_{\beta}$ (粒度集 $P\_R=\{R_1,R_2,\cdots,R_m\}$，$R_j\subseteq R(1\le j\le m)$，$\bigcup_{j=1}^{m}R_j=R$)，因为 $R\subseteq C-\{a\}\subseteq C$，故 $r\left(\sum_{j=1}^{m}R_j,D\right)_{\beta}=r\left(\sum_{l=1}^{o}AC_l,D\right)_{\beta}$ (粒度集 $P\_AC=\{AC_1,AC_2,\cdots,AC_o\}$，$AC_l\subseteq C-\{a\}(1\le l\le o)$，$\bigcup_{l=1}^{o}AC_l=C-\{a\}$)，由定义 7 可得 $r\left(\sum_{l=1}^{o}AC_l,D\right)_{\beta}=r\left(\sum_{i=1}^{n}C_i,D\right)_{\beta}$，即 $a\notin core(C)$，因此可得 $core(C)\subseteq\cap RED(C)$。

2) 设 $a\notin core(C)$，由定义 7 可得 $r\left(\sum_{l=1}^{o}AC_l,D\right)_{\beta}=r\left(\sum_{i=1}^{n}C_i,D\right)_{\beta}$ (粒度集 $P\_AC=\{AC_1,AC_2,\cdots,AC_o\}$，$AC_l\subseteq C-\{a\}(1\le l\le o)$，$\bigcup_{l=1}^{o}AC_l=C-\{a\}$)。$\exists R\subseteq C-\{a\}$，使得 $r\left(\sum_{j=1}^{m}R_j,D\right)_{\beta}=r\left(\sum_{l=1}^{o}AC_l,D\right)_{\beta}$ (粒度集 $P\_R=\{R_1,R_2,\cdots,R_m\}$，$R_j\subseteq R(1\le j\le m)$，$\bigcup_{j=1}^{m}R_j=R$)且 $\forall A\subset R$，$r\left(\sum_{r=1}^{k}A_r,D_{mc}\right)_{\beta}\neq r\left(\sum_{j=1}^{m}R_j,D\right)_{\beta}$ (粒度集 $P\_A=\{A_1,A_2,\cdots,A_k\}$，$A_r\subseteq A(1\le r\le k)$，$\bigcup_{r=1}^{k}A_r=A$)，即 $R\in RED(C-\{a\})$，又因为 $R\subseteq C-\{a\}\subseteq C$，那么 $R\in RED(C)$，因为 $a\notin R$，可得 $a\notin\cap RED(C)$，因此 $\cap RED(C)\subseteq core(C)$。

定义 8 给定决策系统 $DS=(U,AT=C\cup D,V,f)$，粒度集 $P\_B=\{B_1,B_2,\cdots,B_m\}$，$B_i\subseteq B(1\le i\le m)$，$\bigcup_{i=1}^{m}B_i=B$，粒度集 $P\_AB=\{B_{a1},B_{a2},\cdots,B_{am}\}$，$B_l\subseteq B\ (a1\le l\le am)$，$\bigcup_{l=a1}^{am}B_l=B\cup\{b\}$，$B\subseteq B\cup\{b\}\subseteq C$，$\beta\in(0.5,1]$，对 $\forall b\in C-B$ 的外部属性重要度定义为：

$$Sig_{outer}(b,B,D)_{\beta}=r\left(\sum_{l=a1}^{am}B_l,D\right)_{\beta}-r\left(\sum_{i=1}^{m}B_i,D\right)_{\beta}.$$

给定决策系统 $DS=(U,AT=C\cup D,V,f)$，首先通过定义 7 计算 DS 中的核属性集 $core$，每只人工蚂蚁构造解时从 $core$ 开始，从候选属性集中随机选择一个节点 $i\in C-core$，当前人工蚂蚁在节点 i，对 $\forall j\in C-\{core\cup i\}$，$j$ 关于 i 的启发信息定义为：

$$\eta_{ij}=Sig_{outer}(j,\{core\cup i\},D)_{\beta},$$

如果 $\eta_{ij}<\varepsilon$，那么 $\eta_{ij}\leftarrow\varepsilon$，其中 $\varepsilon>0$。

可行解的构造

当前人工蚂蚁在节点 i，依据概率选择下一个节点，概率计算如下：

$$p_{ij}^{k}(t)=\frac{\tau_{ij}^{a}\eta_{ij}^{b}(t)}{\sum_{l\in allowed_k}\tau_{il}^{a}\eta_{il}^{b}(t)},\quad j\in allowed_k .$$

其中 k 表示蚂蚁数；t 表示迭代次数；$allowed_k$ 表示候选属性集；τ_{ij} 表示节点 i 到节点 j 路径上的信息素；η_{ij} 表示节点 j 关于节点 i 的启发信息；$a>0$ 表示信息素相对于启发信息的相对重要性；$b>0$ 表示启发信息相对于信息素的相对重要性。若 $a\gg b$，人工蚂蚁选择下一个节点主要是依据信息素的大小；若 $b\gg a$，人工蚂蚁选择下一个节点主要是依据启发信息的大小。

只要满足以下两个条件之一，人工蚂蚁将停止解的构造：

1) $r\left(\sum_{j=1}^{n}R_j,D\right)_\beta=r\left(\sum_{i=1}^{m}P_i,D\right)_\beta$。其中 R 是蚂蚁构造的当前解（$R_1,R_2,\cdots,R_n\subseteq R$，$\bigcup_{j=1}^{n}R_j=R$；$P_1,P_2,\cdots,P_m\subseteq C$，$\bigcup_{i=1}^{m}P_i=C$）；

2) 当前解的长度 $|R|$ 大于临时最短属性集合的长度。

Table 1. A generalized multi-granularity rough set feature selection algorithm based on ant colony algorithm (GL-AFS)
表 1. 基于蚁群算法的广义多粒度粗糙集特征选择算法

输入：决策系统 $DS=(U,AT=C\cup D,V,f)$，粒度集 $P\_C=\{C_1,C_2,\cdots,C_n\}$，$C_i\subseteq C(1\le i\le n)$，$\bigcup_{i=1}^{n}C_i=C$，最大循环次数 $cycle=100$，$|C|/2$ 只蚂蚁(*Ants*)等一系列参数。

输出：一个属性子集 R 和 R 的长度 L。

步骤 1：初始化 $R=C$，$L=|C|$，迭代次数 $t=0$，$core\leftarrow\varnothing$；

步骤 2：计算 $r\left(\sum_{i=1}^{n}C_i,D\right)_\beta$；

步骤 3：对 $\forall a\in C$，计算 $Sig^{inner}(a,C,D)_\beta$：当 $Sig^{inner}(a,C,D)_\beta\neq 0$ 时，$core\leftarrow core\cup\{a\}$；

步骤 4：当 $t\le cycle$ 时，执行以下操作：

步骤 4.1：对 $k\in Ants$，循环执行以下操作：

步骤 4.1.1：$R_k=core$，$L_k=|core|$；

步骤 4.1.2：随机选择一个属性 $a\in C-core$，$R_k=R_k\cup\{a\}$，$L_k=L_k+1$；

步骤 4.1.3：如果 $r\left(\sum_{j=1}^{m}R_j,D\right)_\beta\neq r\left(\sum_{i=1}^{n}C_i,D\right)_\beta$ 且 $L_k<L$；其中 $R_1,R_2,\cdots,R_m\subseteq R_k$，$\bigcup_{j=1}^{m}R_j=R_k$，则重复：

计算 $p_{ij}^{k}(t)=\max\{p_{ib}^{k}(t),b\in C-R_k\}$，$R_k=R_k\cup\{j\}$，$L_k=L_k+1$；

步骤 4.1.4：如果 $r\left(\sum_{j=1}^{m}R_j,D\right)_\beta=r\left(\sum_{i=1}^{n}C_i,D\right)_\beta$ 且 $L_k<L$：$R=R_k$，$L=|R_k|$；

步骤 4.2：更新信息素：

步骤 4.2.1：对 $\forall x,y\in\{R-core\}$：$\tau_{x,y}=\rho\tau_{x,y}+q/L$；

步骤 4.2.2：对 $\forall u,v\in\{C-R\}$：$\tau_{u,v}=\rho\tau_{u,v}$；

步骤 4.3：$t=t+1$；

步骤 5：返回 R 和 L。

信息素的更新

每轮迭代结束时，可得到一个当前最优解，此时需要对每条路径上的信息素进行更新，信息素依据以下规则更新：

$$\tau_{ij}(t+1)=\rho\tau_{ij}(t)+\Delta\tau_{ij}(t).$$

其中 $\tau_{ij}(t)$ 表示迭代 t 次时路径 (i,j) 上的信息素值； $\tau_{ij}(t+1)$ 表示下一次迭代时路径 (i,j) 上的信息素值； $\rho(0<\rho<1)$ 表示信息素蒸发的衰减常数； $\Delta\tau_{ij}(t)$ 表示路径 (i,j) 上存储的信息素值，计算方式如下：

$$\Delta\tau_{ij}(t)=\begin{cases}q/|R(t)| & (i,j)\text{已遍历}\\ 0 & \text{其他}\end{cases}.$$

其中 q 是给定的常数； $R(t)$ 表示在迭代次数 t 时，当前得到属性集合的长度。

下面将给出通过蚁群算法实现广义多粒度粗糙集特征选择的算法，算法的具体描述见表 1。

表 1 中，步骤 2 用于计算广义多粒度粗糙集的依赖度，步骤 3 用于求取决策系统的核属性集，步骤 4 用于模拟蚂蚁觅食的过程，其中，步骤 4.1.3 为蚂蚁觅食结束的条件，步骤 4.2 为路径上信息素的更新。

4. 实验分析

本节将在运行时间和分类精度两个方面对所提出的算法进行验证。实验选用 6 组标准 UCI 数据集，所用数据集见表 2。数据集通过 WEKA3.6 进行等频离散化，将数据集中名词性数据使用整数进行替换表示。实验所运行的硬件环境为：Windows10 64 位操作系统；8192MB RAM 内存；Intel Core i3-9100 CPU；软件环境为：Pycharm 2020；编程语言：Python。

Table 2. Dataset description
表 2. 数据集描述

编号	数据集	对象数	属性数	类别数
1	OBS-Network-DataSet	1075	21	4
2	Audit_risk	776	26	2
3	Wdbc	569	30	3
4	Congressional Voting Records	435	16	2
5	House	506	13	4
6	Lymph	148	18	4

Table 3. Feature subset length comparison
表 3. 特征子集长度比较

编号	数据集	$\lvert C\rvert$	$num=3$	$num=4$	$num=5$
1	OBS-Network-DataSet	21	15.60	16.55	15.95
2	Audit_risk	26	17.45	18.05	17.35
3	Wdbc	30	10.30	10.10	10.15
4	Congressional Voting Records	16	9.95	15.10	15.25
5	House	13	11.00	11.80	12.05
6	Lymph	18	8.50	15.25	16.35

Table 4. KNN classification accuracy
表 4. KNN 分类精度

编号	数据集	C	$num=3$	$num=4$	$num=5$
1	OBS-Network-DataSet	0.9954	0.9818	0.9942	0.9911
2	Audit_risk	0.9030	0.9057	0.9014	0.8984
3	Wdbc	0.8753	0.8695	0.8690	0.8784
4	Congressional Voting Records	0.7934	0.7807	0.7851	0.7859
5	House	0.6347	0.6270	0.6254	0.6290
6	Lymph	0.7157	0.6971	0.7237	0.7233

本节验证 GL-AFS 算法的有效性，进行了两方面的比较：特征子集长度的比较，见表 3；特征子集分类精度的比较，见表 4、表 5。为了满足多粒度的思想，实验任选 3 个属性看作一个粒度($num=3$)、4 个属性看作一个粒度($num=4$)、5 个属性看作一个粒度($num=5$)，且满足粒度和粒度之间的交集为空，粒度的并集为条件属性集。实验参数：阈值 $\beta=0.6$ 、 $a=1$ 、 $b=0.01$ 、 $\rho=0.9$ 、 $q=0.1$ 、 $\varepsilon=0.001$ 、初始化信息素为 0.5。算法运行停止的条件是：达到最大循环次数或三次迭代过程得到的特征集合相同。由于主要依据信息素进行特征选择、粒度选择的随机性使得蚁群算法每次得到的结果不同，为了保证实验的准确性，将 GL-AFS 算法运行 20 次，取特征选择结果长度的平均值放入表 3 中。通过表 3 可以看出通过本文提出的 GL-AFS 算法可以起到对高维数据集进行降维处理的效果。表 4、表 5 分别为 GL-AFS 算法运行 20 次，对每次得到的结果通过十折交叉验证的方法计算在 KNN、SVM 分类器上的分类精度，分别取 20 次的平均值。通过表 4、表 5 可以看出任选 3 个属性一个粒度、任选 4 个属性一个粒度、任选 5 个属性一个粒度通过 GL-AFS 算法得到特征集合的分类精度和条件属性集 C 下的得到的分类精度的数值相差不大。可以得出，通过 GL-AFS 算法可以得到和原数据分类性能相差不大的特征集合。

Table 5. SVM classification accuracy
表 5. SVM 分类精度

编号	数据集	C	$num=3$	$num=4$	$num=5$
1	OBS-Network-DataSet	0.7628	0.7754	0.7824	0.7791
2	Audit_risk	0.8900	0.9035	0.8886	0.8917
3	Wdbc	0.8594	0.8615	0.8602	0.8660
4	Congressional Voting Records	0.8529	0.8540	0.8533	0.8538
5	House	0.7529	0.7242	0.7263	0.7345
6	Lymph	0.7838	0.7414	0.7832	0.7808

5. 总结

目前针对广义多粒度粗糙集特征选择的研究不完善，通过元启发式算法进行广义多粒度粗糙集特征选择未有人研究，因此，本文将元启发式算法(蚁群算法)用于广义多粒度粗糙集特征选择中具有很重要的研究意义。实验表明：本文所提算法不仅可以对高维数据实现降维的效果且得到的特征集合具有和原数据集相差不大的分类精度。

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] Qian, Y.H., Liang, J.Y., Yao, Y.Y. and Dang, C.Y. (2009) MGRS: A Multi-Granulation Rough Set. *Information Sciences*, **180**, 949-970. https://doi.org/10.1016/j.ins.2009.11.023

[3] Xu, W.H., Zhang, X.T. and Wang, Q.R. (2012) A Generalized Multi-Granulation Rough Set Approach. *International Conference on Intelligent Computing*, Zhengzhou, 11-14 August 2011, 681-689. https://doi.org/10.1007/978-3-642-24553-4_90

[4] Qian, J., Hong, C.X., Yu, Y., Liu, C.H. and Miao, D.Q. (2022) Generalized Multigranulation Sequential Three-Way Decision Models for Hierarchical Classification. *Information Sciences*, **616**, 66-87. https://doi.org/10.1016/j.ins.2022.10.014

[5] Xu, W.H. and Guo, Y.T. (2016) Generalized Multigranulation Double-Quantitative Decision Theoretic Rough Set. *Knowledge Based Systems*, **105**, 190-205. https://doi.org/10.1016/j.knosys.2016.05.021

[6] Xu, W.H., Yuan, K.H. and Li, W.T. (2022) Dynamic Updating Approximations of Local Generalized Multigranulation Neighborhood Rough Set. *Applied Intelligence*, **52**, 9148-9173. https://doi.org/10.1007/s10489-021-02861-x

[7] 张先韬. 广义多粒度粗糙集属性约简和 matlab 计算[J]. 计算机工程与应用, 2016, 52(8): 43-48.

[8] Xu, W.H., Li, W.T. and Zhang, X.T. (2017) Generalized Multigranulation Rough Sets and Optimal Granularity Selection. *Granular Computing*, **2**, 271-288. https://doi.org/10.1007/s41066-017-0042-9

[9] Aram, K.Y., Lam, S.S. and Khasawneh, M.T. (2023) Cost-Sensitive Max-Margin Feature Selection for SVM Using Alternated Sorting Method Genetic Algorithm. *Knowledge-Based Systems*, **267**, Article ID: 110421. https://doi.org/10.1016/j.knosys.2023.110421

[10] Zhong, C.T., Li, G., Meng, Z., Li, H.J. and He, W.X. (2023) A Self-Adaptive Quantum Equilibrium Optimizer with Artificial Bee Colony for Feature Selection. *Computers in Biology and Medicine*, **153**, Article ID: 106520. https://doi.org/10.1016/j.compbiomed.2022.106520

[11] Dorigo, M. and Caro, G.D. (1999) Ant Colony Optimization: A New Meta-Heuristic. *Congress on Evolutionary Computation* (*CEC*99), Vol. 2, 1470-1477. https://doi.org/10.1109/CEC.1999.782657

[12] Jensen, R. and Shen, Q. (2013) Finding Rough Set Reducts with Ant Colony Optimization. *Proceedings of the UK Workshop on Computational Intelligence*, **1**, 15-22.

[13] Chen, Y.M., Miao, D.Q. and Wang, R.Z. (2010) A Rough Set Approach to Feature Selection Based on Ant Colony Optimization. *Pattern Recognition Letters*, **31**, 226-233. https://doi.org/10.1016/j.patrec.2009.10.013