计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181202285

基于粗糙集和改进鲸鱼优化算法的特征选择方法

王生武　陈红梅
西南交通大学信息科学与技术学院　成都 611756
西南交通大学云计算与智能技术高校重点实验室　成都 611756
(shengwuwang@my.swjtu.edu.cn)

摘　要　随着互联网和物联网技术的发展,数据的收集变得越发容易。但是,高维数据中包含了很多冗余和不相关的特征,直接使用会徒增模型的计算量,甚至会降低模型的表现性能,故很有必要对高维数据进行降维处理。特征选择可以通过减少特征维度来降低计算开销和去除冗余特征,以提高机器学习模型的性能,并保留了数据的原始特征,具有良好的可解释性。特征选择已经成为机器学习领域中重要的数据预处理步骤之一。粗糙集理论是一种可用于特征选择的有效方法,它可以通过去除冗余信息来保留原始特征的特性。然而,由于计算所有的特征子集组合的开销较大,传统的基于粗糙集的特征选择方法很难找到全局最优的特征子集。针对上述问题,文中提出了一种基于粗糙集和改进鲸鱼优化算法的特征选择方法。为避免鲸鱼算法陷入局部优化,文中提出了种群优化和扰动策略的改进鲸鱼算法。该算法首先随机初始化一系列特征子集,然后用基于粗糙集属性依赖度的目标函数来评价各子集的优劣,最后使用改进鲸鱼优化算法,通过不断迭代找到可接受的近似最优特征子集。在UCI数据集上的实验结果表明,当以支持向量机为评价所用的分类器时,文中提出的算法能找到具有较少信息损失的特征子集,且具有较高的分类精度。因此,所提算法在特征选择方面具有一定的优势。

关键词:特征选择;粗糙集理论;改进鲸鱼优化算法;属性依赖度;最优特征子集
中图法分类号　TP301.6

Feature Selection Method Based on Rough Sets and Improved Whale Optimization Algorithm

WANG Sheng-wu and CHEN Hong-mei
School of Information Science and Technology,Southwest Jiaotong University,Chengdu 611756,China
Key Laboratory of Cloud Computing and Intelligent Technology,Southwest Jiaotong University,Chengdu 611756,China

Abstract　With the development of the Internet and Internet of Things technologies,data collection has become easier. However,it is necessary to reduce the dimensionality of high-dimensional data. High-dimensional data contain many redundant and unrelated features,which will increase the computational complexity of the model and even reduce the performance of the model. Feature selection can reduce the computational cost and remove redundant features by reducing feature dimensions to improve the performance of a machine learning model,and retain the original features of the data,with good interpretability. It has become one of important data preprocessing steps in machine learning. Rough set theory is an effective method which can be used to feature selection. It preserves the characteristics of the original features by removing redundant information. However,it is difficult to find the global optimal feature subset by using the traditional rough sets-based feature selection method because the cost of computing all feature subset combinations is very high. In order to overcome above problems,a feature selection method based on rough sets and improved whale optimization algorithm was proposed. An improved whale optimization algorithm was proposed by employing politics of population optimization and disturbance so as to avoid local optimization. The algorithm first randomly initializes a series of feature subsets,and then uses the objective function based on the rough sets attribute dependency to evaluate the goodness of each subset. Finally,the improved whale optimization algorithm is used to find an acceptable approximate optimal feature subset by iterations. The experimental results on the UCI dataset show that the proposed algorithm can find a subset of features with less information loss and has higher classification accuracy when the support vector machine is used as the classifier for evaluation. Therefore,the proposed algorithm has a certain advantage in feature selection.

Keywords　Feature selection,Rough set theory,Improved whale optimization algorithm,Attribute dependency,Optimal feature subset

1　引言

随着数字图像、金融时间序列和基因表达微阵列等高维数据的快速积累,特征选择已成为机器学习和数据挖掘的重要预处理步骤[1]。特征选择方法旨在选择能够改进学习模型性能的特征[2]。此外,特征选择通过去除不相关、冗余的或噪

到稿日期:2018-12-10　返修日期:2019-05-01　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61572406)
This work was supported by the National Natural Science Foundation of China (61572406).
通信作者:陈红梅(hmchen@swjtu.edu.cn)

声的特征属性来减少数据的维数，得到具有更好泛化能力且更加紧凑的模型，进而降低数据存储和处理的成本[3-4]。

根据评价准则的不同，现有特征选择算法可分为过滤式(Filter)方法、封装式(Wrapper)方法和嵌入式(Embedded)方法[5-6]。过滤式方法通过使用训练数据的固有性质来评估特征的好坏，并且独立于任何学习算法，有较好的通用性和较高的搜索效率[7]。与此相反，封装式方法直接使用预定义的学习算法来评价特征。由于评价准则与算法的学习目的一致，因此将选出的特征用于预定义的学习算法会具有很高的性能，但同时也极大地增加了计算复杂度[8]。嵌入式方法旨在评估学习阶段的最佳特征子集，通常与特定的分类任务绑定，不具有通用性[9]。一般而言，在已知后续学习算法的前提下，封装式方法选出的特征的效果明显优于过滤式方法选出的特征的效果，但前者的计算开销远大于后者。在处理高维大数据集时，庞大的计算开销使得封装式方法显得不适用，而过滤式方法被经常采用以节省计算时间。

Pawlak 提出的粗糙集理论(Rough Set Theory，RST)是处理不确定、不准确和模糊数据的有效数学工具[10]。RST 是能有效进行特征选择的方法，能通过删除冗余特征来保留原始特征的性质。RST 已被广泛应用于分类、特征选择、模式识别、图像处理、离群检测和信息检索等[11-16]领域。

用粗糙集进行特征选择可以得到不同的约简结果，通常以减小计算开销和提高分类器的性能为准则，定义具有最少特征个数并具有较好分类性能的属性子集为较优的约简结果。Hu 等给出了 Pawlak 粗糙集模型下的优化算法 FARCF(Forward Attribute Reduction Based on Classical Rough Sets and Fast Search)，该算法在选择添加新的特征时，着重寻找能够使边界域缩减最快的特征，极大地提高了属性约减的速度[17]。然而，最小约简子集的产生已被证明是 NP-难问题，并且产生所有约简子集所需要的计算时间是指数级的[18]。因此，考虑将计算开销低且全局寻优能力强的群智能算法与粗糙集相结合来进行特征选择，是一种行之有效的方法。

群智能算法具有概念简单、易于实现、全局寻优能力强等特点，因此被广泛用于维度约简，如基因编程(Genetic Programming，GP)[19]、基因算法(Genetic Algorithms，GAs)[20]、和声搜索(Harmony Search，HS)算法[21]、果蝇优化算法(Fruit Fly Optimization Algorithm，FOA)[22]、蚁群优化(Ant Colony Optimization，ACO)算法[23]、粒子群优化(Particle Swarm Optimization，PSO)算法[24]等。但大部分基于群智能算法的特征选择方法都是封装式方法，高昂的计算开销使得其在面对现实生活中的高维数据时很难奏效。而粗糙集的属性约简方法为过滤式特征选择方法提供了一种崭新的思路。粗糙集能够去除不相关、冗余的特征，同时保留原始特征的特性，但具有 NP-难的问题，很难找到最优解，而这一问题正好可以用群智能算法来弥补。Wang 等早些时期就提出了一种新颖的二进制粒子群的位置更新方法，并将其与粗糙集的属性依赖度相结合，得到新的特征选择方法 PSORSFS(Particle Swarm Optimization and Rough Set - based Feature Selection)[25]。Wang 等提出了一种基于粗糙集和蚁群优化算法的特征选择方法，该算法根据特征子集长度和属性依赖度构建信息素更新策略，在较小数据集下具有较好的性能，但随着数据集的增大，容易陷入局部最优，算法效率下降[26]。Chen 等提出了 FSARSR(Fish Swarm Algorithm for Rough Set Reduction)算法，重新定义了鱼群算法的距离和中心点的计算方式，并将其与粗糙集相结合进行属性约简[27]。Mirjalili 等受座头鲸捕食行为的启发，提出一种基于自然灵感的新型群智能算法——鲸鱼优化算法(Whale Optimization Algorithm，WOA)[28]。该算法模仿座头鲸利用“螺旋气泡网”策略，并通过收缩包围、螺旋式位置更新及随机捕猎机制进行觅食，其结构简单、收敛速度快；更为重要的是，与其他传统的优化算法相比，鲸鱼优化算法具有更强的全局寻优能力。本文将鲸鱼优化算法离散化，并将其加入群体划分策略，在精英群体中加入局部扰动，使鲸鱼个体在靠近最优个体时具有更强的搜索能力，避免过早陷入局部最优，再将粗糙集中属性依赖度的概念用于适应度函数中，提出了一种新颖的特征选择方法。

2 相关基础知识

本节将介绍粗糙集理论和鲸鱼优化算法的相关知识。

2.1 粗糙集的相关原理

粗糙集(Rough Set，RS)是处理不精确、不确定和不完全数据的有效数学工具。RS 的研究对象是决策表(信息表)，且不需要额外的先验知识来分析数据。

定义 1 信息系统 S 可由一个四元组 $S=(U,A,V,f)$ 表示，其中：

1) $U=\{x_1,x_2,\cdots,x_{|U|}\}$ 是非空有限的对象集合，称为论域；

2) $A=\{a_1,a_2,\cdots,a_{|A|}\}$ 是非空有限的属性集合；

3) $V=\bigcup_{a\in A}V_a$ 是值域，V_a 表示在论域 U 中属性 a 相同的一组对象；

4) $f:U\times A_a\to V_a$ 是信息函数，即 $\forall x\in U,a\in A,f(x,a)\in V_a$。

如果 $A=C\cup D$，其中 C 是条件属性，D 是决策属性，则称 $S=(U,A,V,f)$ 为一个决策表。

定义 2 给定信息系统 $S=(U,A,V,f)$，对于任意条件的属性子集 $B\subseteq A$，论域 U 上的一个不可分辨关系定义为：

$$IND(B)=\{(x,y)\in U\times U\mid \forall b\in B,f(x,b)=f(y,b)\} \tag{1}$$

其中，$f(x,b)$ 和 $f(y,b)$ 分别表示对象 x 和 y 在条件属性 b 上的属性值。该不可分辨关系为一个等价关系，通常记为 R_B。

等价关系 R_B 在论域 U 上形成了一个划分 $\frac{U}{R_B}$，即等价类集合 $\frac{U}{R_B}=\{[x]_B\mid x\in U\}$，其中 $[x]_B$ 表示对象集合 U 中由等价关系 R_B 所确定的等价类，即 $[x]_B=\{y\in U\mid (x,y)\in R_B\}$。在不引起歧义的情况下，$\frac{U}{R_B}$ 可简写为 $\frac{U}{B}$。

定义 3 给定信息系统 $S=(U,A,V,f)$，对于任意对象集合 $X\subseteq U$ 和 U 上的一个等价关系 B，X 关于 B 的下、上近似集分别定义为：

$$\underline{B}(X)=\{x\in U\mid [x]_B\subseteq X\} \tag{2}$$

$$\overline{B}(X)=\{x\in U\mid [x]_B\cap X\neq\varnothing\} \tag{3}$$

定义 4 给定决策表 $S=(U,C\cup D,V,f)$，C 和 D 分别表示条件属性和决策属性，设 $B\subseteq C$ 为条件属性的任一子集，则决策属性 D 在条件属性 B 下的正域定义为：

$$POS_B(D)=\bigcup_{X\in U/D}\underline{B}(X) \tag{4}$$

其中，$POS_B(D)$ 表示根据 B 的知识进行划分得到 $\frac{U}{B}$ 后能确切

划入$\frac{U}{D}$的对象集合。

定义 5 给定决策表 $S=(U,C\cup D,V,f)$，$C\cup D$ 分别表示条件属性和决策属性，设 $B\subseteq C$ 为条件属性的任意子集，则称 B 为 C 的一个约简，如果它满足：

1) $POS_B(D)=POS_C(D)$；

2) $\forall b\in B, POS_{B-\{b\}}(D)\neq POS_B(D)$。

定义 6 决策属性 D 对条件属性 C 的依赖度 $\gamma_c(D)$ 定义为：

$$\gamma_c(D)=\frac{|POS_c(D)|}{|U|} \tag{5}$$

依赖度 $\gamma_c(D)$ 表示在条件属性 C 下，确切能划入决策 $\frac{U}{D}$ 的对象在整个论域中所占的比重。

2.2 鲸鱼优化算法

鲸鱼优化算法是受座头鲸捕食行为启发所提出的一种新型启发式算法。座头鲸有着与众不同的捕猎方式——气泡网觅食，通过不断地收缩包围、螺旋式地更新位置和随机捕猎的机制进行觅食。下文介绍座头鲸觅食过程中的 3 个行为。

2.2.1 环绕猎物

座头鲸可以识别猎物的位置，并将它们包围起来。由于搜索空间中的最优个体的位置不是先验已知的，因此 WOA 算法假设当前适应度最高的个体(最佳个体)的位置是目标位置或接近最优的位置。在定义最佳个体之后，其他个体将尝试朝着最优个体的位置更新自己的位置，此行为的表达式为：

$$\vec{D}=|C\cdot\vec{X}^*(t)-\vec{X}(t)| \tag{6}$$

$$\vec{X}(t+1)=\vec{X}^*(t)-A\cdot\vec{D} \tag{7}$$

其中，t 表示当前迭代次数，A 和 C 是系数变量，$\vec{X}^*$ 是当前最佳个体的位置向量，$\vec{X}$ 是位置向量，$\|$ 是取绝对值运算符，$\cdot$ 是逐个元素的乘法。系数变量 A 和 C 的计算公式如下：

$$A=2a\cdot r_1-a \tag{8}$$

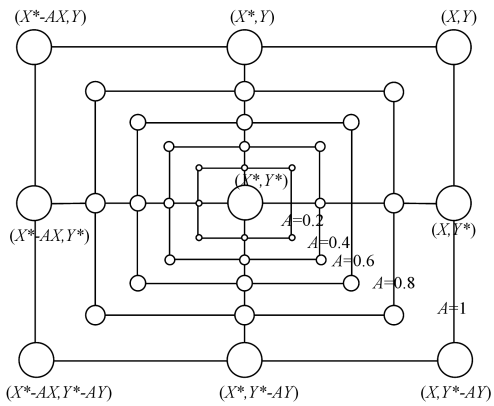
$$C=2r_2 \tag{9}$$

其中，a 是一个随着迭代开始从 2 衰减到 0 的系数，r_1 和 r_2 是 $[0,1]$之间的随机变量。

2.2.2 气泡网攻击

鲸鱼优化算法设计了两种方法在数学模型上模拟座头鲸的气泡网行为。

1)收缩环绕机制。该行为是通过减小式(7)中的系数 A 来实现的。



注：X^* 表示当前最优个体的位置

图 1 收缩环绕机制

Fig. 1 Shrinking encircling mechanism

由式(8)可以看出，系数 A 的波动范围是受系数 a 影响的，实际为$[-a,a]$之间的随机值。随着系数 a 从 2 衰减到 0，个体更新的位置也越来越贴近当前最佳个体的位置。图 1 展示了当$0\leqslant A\leqslant 1$ 时，鲸鱼个体向最优个体收缩环绕的过程中可能出现的位置。

2)螺旋更新位置。该方法首先计算个体位置到最优个体位置的距离，然后使用螺旋线公式来模拟座头鲸的螺旋环绕行为，如式(10)所示：

$$\vec{X}(t+1)=\vec{D'}\cdot e^{bl}\cdot\cos(2\pi l)+\vec{X}^*(t) \tag{10}$$

其中，$\vec{D'}=|\vec{X}^*(t)-\vec{X}(t)|$，表示个体位置与最优个体位置之间的距离；$b$ 是定义螺旋线形状的常量；l 是$[-1,1]$之间的随机数。

座头鲸在螺旋环绕猎物的同时收缩包围圈。为了用数学模型模拟同时进行这两种行为，假设分别有 50%的概率来选择收缩环绕行为和螺旋更新行为。气泡网攻击的数学模型如下：

$$\vec{X}(t+1)=\begin{cases}\vec{X}^*(t)-A\cdot\vec{D}, & \text{if } p<0.5\\ \vec{D'}\cdot e^{bl}\cdot\cos(2\pi l)+\vec{X}^*(t), & \text{if } p\geqslant 0.5\end{cases} \tag{11}$$

其中，p 是$[0,1]$之间的随机数。

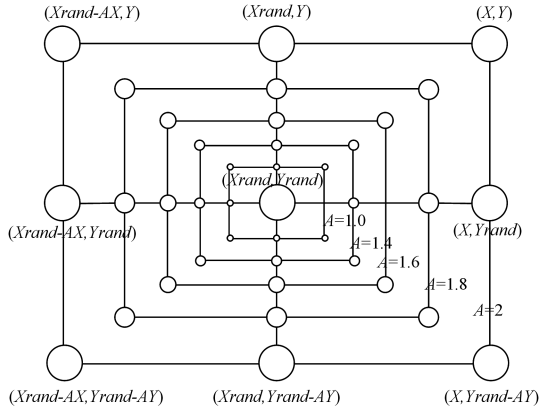
2.2.3 随机更新位置

同样地，可以利用变量 A 的变化来搜索猎物。事实上，座头鲸根据彼此的位置进行随机搜索。因此，设定当$|A|>1$ 时，强制座头鲸根据随机选择的个体而不是最优个体来更新自己的位置，这个机制使得 WOA 算法有着更强的全局搜索能力。数学模型如下：

$$\vec{D}=|C\cdot\vec{X}_{rand}-X| \tag{12}$$

$$\vec{X}(t+1)=\vec{X}_{rand}-A\cdot\vec{D} \tag{13}$$

其中，$\vec{X}_{rand}$ 是从当前群体中随机选择的个体。当$|A|>1$ 时，个体可能的位置如图 2 所示。



注：$Xrand$ 表示随机个体的位置

图 2 随机搜索机制

Fig. 2 Random exploration mechanism

3 基于粗糙集和鲸鱼优化算法的特征选择方法

用粗糙集直接进行特征选择会遇到 NP-难问题，此时需要有一定的策略从众多的候选解中快速、明确地找到最优解，

以节省计算量。而鲸鱼算法提供了一个从随机初始候选解到近似最优候选解的选择路线，且与其他优化算法相比，鲸鱼优化算法的随机搜索策略使得其具有更强的全局寻优能力，能够以更大的概率得到最优解。本文进一步加强了鲸鱼算法的全局寻优能力，因此将两者结合来进行特征选择即可较好地解决粗糙集的 NP-难问题。此外，通过定义一个良好的评价函数，可以对特征子集做进一步的筛选，从而得到更符合预期的特征子集。基于此，本文提出一种基于粗糙集与改进鲸鱼优化算法的特征选择方法（Feature Selection Method Based on Rough Sets and Improved Whale Algorithm，FSRSWOA）。在将鲸鱼优化算法用于特征选择之前，有两个重要的问题需要解决：1）原始的鲸鱼优化算法用于连续运动空间，必须要有一定的方法使得鲸鱼优化算法能够进行离散的特征选择优化；2）适应度函数的定义，合适的适应度函数可以帮助优化算法选出性能最优、最符合需求的特征子集。此外，本文对鲸鱼优化算法的全局寻优进行了改进，引入了群体划分策略，在每次迭代后将鲸鱼群体整体划分为精英群体和普通群体，并在精英群体中引入扰动的思想来加强算法的局部搜索能力。

3.1 鲸鱼优化算法的离散化方法

设鲸鱼个体的位置为 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$，在算法中将鲸鱼优化算法中个体的运动范围限制在$[0,1]$，即 $x_i\in[0,1]$，则鲸鱼个体的位置由一组$[0,1]$之间的小数组成。根据一定的转化方法，可将连续运动空间中的个体位置 X 转化为长度为 n 的二进制位置 $X'=\{x_1',x_2',\cdots,x_i',\cdots,x_n'\}$，其中 $x_i'\in\{0,1\}$，n 是特征的个数。每一个比特位 x_i'代表一个特征，值为 1 表示该特征被选择，值为 0 表示该特征未被选择。每一个二进制位置都表示一个特征子集。鲸鱼的二进制位置 X'可由式(14)得到：

$$x_i'=\begin{cases}1, & \text{if } rand<x_i\\ 0, & \text{otherwise}\end{cases} \tag{14}$$

其中，x_i 表示鲸鱼个体在连续运动空间中的位置 X 的第 i 维数值，$rand$ 表示$[0,1]$之间的随机数，这样可以使得搜索范围更大，全局寻优能力更强。

3.2 适应度函数

合适的适应度函数对特征选择方法的研究具有重要的作用。本文采用的适应度函数如下：

$$Fitness=\alpha\cdot\gamma_R(D)+\beta\cdot\frac{|C|-|R|}{|C|} \tag{15}$$

其中，$|R|$为当前所选特征子集的长度，$|C|$为数据集特征属性的总个数，$\gamma_R(D)$为条件属性 R 相对于决策属性 D 的依赖度。α 和 β 是分别对应于属性依赖度和特征子集长度重要性的两个参数，$\alpha\in[0,1]$，$\beta=1-\alpha$。该公式意味着属性依赖度和子集长度对特征选择任务有着不同的意义。在实验中，我们认为属性依赖度比特征长度更加重要，并设置 $\alpha=0.9$，$\beta=0.1$，因为属性依赖度是正域所占比例，而正域可以导出确定性规则，使属性依赖度占据高的比例主要是为了保证分类精度。通过该适应度函数来评价各特征子集的优劣，适应度值越大的特征子集越优。

3.3 群体划分策略

设当前鲸鱼群体为 $POP(t)$，利用式(15)计算群体中各鲸鱼个体的适应度 $Fitness_i$，并得到 $POP(t)$中的最佳适应度值 $Fitness_{\text{best}}$和最差适应度值 $Fitness_{\text{worst}}$。用 $d_{i,j}=|Fitness_i-Fitness_j|$分别计算鲸鱼个体 i 与最优个体的适应度差值 $d_{i,\text{best}}$和与最差个体的适应度差值 $d_{i,\text{worst}}$。若 $d_{i,\text{best}}$和 $d_{i,\text{worst}}$满足 $d_{i,\text{best}}\leqslant\exp(-\frac{Iter}{Iter_{\text{best}}})d_{i,\text{worst}}$（其中，$Iter$ 为当前迭代次数，$Iter_{\max}$为最大迭代次数），则鲸鱼个体 i 被划分到精英子群 $POP_e(t)$，否则被划分到普通子群 $POP_g(t)$。

3.4 局部扰动

对于普通子群 $POP_g(t)$中的鲸鱼个体而言，其距离当前最优个体还较远，在向当前最优个体靠近的过程中还有很多可能，不太容易陷入局部最优；但精英子群 $POP_e(t)$中的鲸鱼个体距离当前最优个体很近，很容易快速收敛，陷入局部最优。

针对上述情况，根据式(16)构造最优个体替代点$\vec{X}_m$：

$$\vec{X}_m=\vec{X}^*-m\cdot c\cdot|\vec{X}^*-\vec{X}_{rand'}| \tag{16}$$

其中，$m=\frac{Iter_{\max}-Iter}{Iter_{\max}}$，$c$ 是$[0,1]$之间的随机数，$\vec{X}^*$ 是当前最优个体，$\vec{X}_{rand'}$是随机生成的个体。

对于精英子群中的鲸鱼个体而言，其位置更新过程将使用$\vec{X}_m$ 来代替式(11)中的最优个体$\vec{X}^*$，而普通子群中的鲸鱼个体的位置更新过程保持不变。

3.5 基于粗糙集和改进鲸鱼优化算法的特征选择方法

FSRSWOA 算法的核心思想是用二进制的鲸鱼位置来表示一个特征子集的选择结果，然后使用基于粗糙集属性依赖度的适应度函数来评价该位置的优劣，不断朝着当前最优个体所在的方向移动自己的位置，最后利用改进的鲸鱼优化算法的全局寻优能力找到特征空间中的最优解。

FSRSWOA 算法的大致过程可分为以下几个步骤：

1）随机产生 N 个各维数值在$[0,1]$范围内的鲸鱼群体；

2）根据式(14)将各鲸鱼的位置转化成相应的 0-1 向量；

3）利用 3.2 节给出的适应度函数计算各鲸鱼的适应度值；

4）根据适应度值对群体进行划分，对于精英子群，使用最优个体替代点来进行位置更新；

5）根据系数变量 A 的值进行相应的捕食行为，并更新鲸鱼个体的位置；

6）若达到最大迭代次数，则算法终止，否则转入步骤 2）。

以上过程大致描述了本文提出的使用鲸鱼算法结合粗糙集理论进行特征选择的框架。FSRSWOA 算法的详细描述如算法 1 所示。

算法 1 FSRSWOA 算法

Input：决策信息系统 S=(U,C∪D,V,f)，鲸鱼群体大小 N，最大迭代次数 $\text{Iter}_{\max}$，权重系数 α 和 β

Output：特征子集 R⊆C

1. 随机产生大小为 N、维度为|C|的鲸鱼群体 $\text{POP}=\{\text{X}_1,\text{X}_2,\cdots,\text{X}_i,\cdots,\text{X}_N\}$，其中 $\text{X}_i=\{\text{x}_1,\text{x}_2,\cdots,\text{x}_i,\cdots,\text{x}_{|C|}\}$，$\text{x}_i\in[0,1]$。
2. For iter=1,2,…，$\text{Iter}_{\max}$
3. 根据式(14)将数值鲸鱼群体 POP(iter)转化为二进制鲸鱼群体 $\text{POP}'(\text{iter})=\{\text{X}_1',\text{X}_2',\cdots,\text{X}_i',\cdots,\text{X}_N'\}$。
4. 将 POP′(iter)代入式(5)计算各鲸鱼个体的属性依赖度 $\gamma_{\text{X}_i'}(\text{D})$。
5. 根据式(15)计算各鲸鱼个体的适应度值 Fitness_i。
6. 根据适应度值将群体 POP′(iter)划分为精英子群 $\text{POP}_e'(\text{iter})$和普通子群 $\text{POP}_g'(\text{iter})$。
7. 根据式(16)计算最优替代点$\vec{\text{X}}_m$。

8. If $X_i \in POP_e'(iter)$,则
9. 使用$\vec{X}_m$代替后续公式中的$\vec{X}^*$;
10. If $|A| < 1$,则
11. If rand > 0.5,则
12. 根据式(7)更新各鲸鱼个体的位置 X_i;
13. Else
14. 根据式(11)更新各鲸鱼个体的位置 X_i;
15. Else
16. 根据式(13)更新各鲸鱼个体的位置 X_i。
17. End For

4 实验设计与结果分析

为了验证本文算法的有效性,在来自 UCI 的真实数据集上用 FSRSWOA 算法对其进行特征选择。使用分类器计算选择出来的结果的其分类精度。通过对实验结果的分析证明,FSRSWOA 算法在大部分情况下能够选出简短且具有高分类质量的特征子集。

4.1 数据集

实验所用数据集来自 UCI 数据库,为验证本文算法的有效性,选取了规模大小各异、特征数目不同的 10 个数据集进行实验。表 1 列出了这些数据集的基本信息,包括数据集名称、样本个数、数值属性个数、符号属性个数及类别数。其中各数据集的基本信息可以看到,部分数据集(如 wine, spectfeart, vehicle 和 bands)中包含了数值属性,然而基于粗糙集的特征选择方法无法直接运用到包含数值属性的数据集上。由于现实生活中包含数值属性的数据普遍存在,为了能够将基于粗糙集的特征选择方法用到包含数值属性的数据集上,需要在实验开始之前对数据集中的数值数据进行离散化处理。目前,对数值数据进行离散化的方法有很多,如等宽离散化、等频离散化和基于熵的离散化方法等。本文使用的是基于卡方的离散化方法,用 R 语言中 discretization 包中的 chiM 函数实现。

表 1 数据集

Table 1 Date set

名称	样本个数	数值属性个数	符号属性个数	类别数
zoo	101	0	16	7
wine	178	13	0	3
spectfheart	267	44	0	2
vehicle	846	18	0	4
bands	365	19	0	2
flare	1066	0	11	6
tic-tac-toe	958	0	9	2
vote	534	0	16	2
breast-cancer	286	0	9	2
mushroom	5644	0	22	2

4.2 分类精度实验及分析

基于表 1 给出的 10 个数据集,使用 FSRSWOA 算法及其对比算法进行特征选择实验,采用的对比算法为 FARCF 算法[17]、PSORSFS 算法[25]以及 FSARSR 算法[27]。实验中,种群大小设为 30,迭代次数为 100。由于基于群智能算法的特征选择都具有一定的随机性,为避免偶然性误差对实验结果产生的影响,FSRSWOA, FSARSR 及 PSORSFS 算法在每个数据集上均进行了 10 次独立实验。同时,本文使用的分类算法为 weka[29]中的 libSVM 算法,采用十折交叉验证的方法来测定分类精度。由于十折交叉验证的结果与数据集的划分密切相关,为避免极端情况下的划分对分类结果产生影响,对各算法的特征选择结果均进行了 20 次十折交叉验证实验,交叉验证过程中的数据集都随机进行划分。

表 2 列出各算法在各数据集上最终选出的特征结果。FARCF 算法是 Hu 等对经典粗糙集算法在速度上进行改进的结果,该算法每次都挑选能够使得正域增长最快的特征加入特征子集,在高维数据上其时间开销是让人难以接受的;但对于低维数据,在时间开销可以接受的情况下,它总能选出属性依赖度最高且特征子集最短的结果。只有在部分情况下,在适应度函数式(15)中特征长度限制的影响下,基于优化算法和粗糙集的特征选择算法才会选出更短的结果,否则都与 FARCF 算法选出的特征长度接近。以此为参照,从表 2 中可以看出,在大部分情况下,与 FSARSR 算法和 PSOREFS 算法相比,FSRSWOA 算法都能选出最佳的结果;且其他不是最短特征子集的情况也与最短特征子集的长度极为接近。

表 2 特征子集大小的比较

Table 2 Comparison of number of feature subsets

数据集	FARCF	FSARSR	PSORSFS	FSRSWOA
zoo	5	$5^{(1)}6^{(9)}$	$8^{(5)}9^{(5)}$	$7^{(2)}8^{(5)}9^{(3)}$
wine	3	$3^{(6)}4^{(4)}$	$3^{(2)}4^{(6)}5^{(2)}$	$3^{(6)}4^{(4)}$
spectfheart	6	$11^{(1)}12^{(5)}14^{(4)}$	$11^{(1)}12^{(1)}13^{(1)}$ $14^{(2)}15^{(3)}16^{(2)}$	$8^{(2)}9^{(2)}$ $10^{(4)}11^{(2)}$
vehicle	7	$8^{(8)}9^{(2)}$	$8^{(3)}9^{(6)}10^{(1)}$	$7^{(3)}8^{(5)}9^{(2)}$
bands	6	$7^{(10)}$	$7^{(3)}8^{(3)}9^{(4)}$	$6^{(3)}7^{(1)}8^{(6)}$
flare	10	$10^{(10)}$	$6^{(1)}7^{(1)}8^{(2)}$ $9^{(4)}10^{(2)}$	$9^{(3)}10^{(7)}$
tic-tac-toe	—	$8^{(10)}$	$7^{(6)}8^{(4)}$	$7^{(6)}8^{(4)}$
vote	—	$9^{(2)}10^{(7)}11^{(1)}$	$8^{(3)}9^{(4)}$ $10^{(2)}11^{(1)}$	$7^{(1)}8^{(1)}$ $9^{(2)}10^{(6)}$
breast-cancer	9	$8^{(10)}$	$6^{(2)}7^{(5)}8^{(2)}9^{(1)}$	$7^{(4)}8^{(6)}$
mushroom	3	$4^{(7)}5^{(3)}$	$6^{(1)}7^{(4)}8^{(2)}$ $9^{(2)}10^{(1)}$	$3^{(5)}4^{(3)}5^{(2)}$

传统的基于粗糙集的特征选择方法都是单独计算基于各特征的正域,排序选取正域增长最大的特征,一旦遇到各特征的所有取值在所有类别上均有分布,则基于单个特征算出的正域均为 0,算法失效。而基于演化算法和粗糙集的特征选择方法则没有这方面的问题,初始化的群体提供了多种特征选择方案,适应度低的特征选择方案会被淘汰。

表 3 列出各算法在 10 个数据集上用 libSVM 分类器得出的分类精度。其中,FARCF 算法进行了 20 次十折交叉验证,共有 200 个分类结果;而 FSARSR, PSORSFS 和 FSRSWOA 算法都进行了 10 次独立实验,即都进行了 200 次十折交叉验证,分别有 2000 个分类结果。在数目众多的分类结果上进行实验分析,可以极大程度地杜绝偶然性因素对实验结果产生的影响。实验挑选各算法所得所有分类精度的最高值与均值进行分析。从表 3 中的数据可以看到,就最佳分类精度而言,FSRSWOA 算法除了在数据集 wine 上略逊于 FSARSR 算法和在数据集 tic-tac-toe 上略逊与 PSORSFS 外,在其他数据集上,其均高于 FARCF, FSARSR 和 PSORSFS 算法;就平均分类精度而言,FSRSWOA 算法除了在数据集 bands, flare 和 mushroom 上略逊于 FARCF 外,在其他数据

集上，其平均分类精度均高于 FARCF，FSARSR 和 PSORSFS 算法。

表 2 和表 3 中的“—”表示 FARCF 算法无法在 tic-tac-toe 和 vote 数据集上选出结果。

表 3 特征子集分类精度的比较

Table 3 Comparison of classification accuracy of feature subsets

数据集	FARCF		FSARSR		PSORSFS		FSRSWOA	
	Best	Avg	Best	Avg	Best	Avg	Best	Avg
zoo	0.8727	0.8586	**0.9127**	0.8729	0.9118	0.8733	0.9118	**0.8742**
wine	0.5395	0.5143	0.9775	0.8924	0.9716	0.8983	**0.9778**	**0.9121**
spectfheart	0.7944	0.7940	0.7953	0.7941	0.7977	0.7940	**0.8023**	**0.7944**
vehicle	0.2778	0.2519	0.6550	0.6079	0.6844	0.6160	**0.6972**	**0.6362**
bands	0.6471	**0.6373**	0.6354	0.6303	0.6525	0.6339	**0.6440**	0.6325
flare	0.7505	0.7465	0.7533	0.7454	0.7524	0.7460	**0.7533**	**0.7468**
tic-tac-toe	—	—	0.8299	0.7776	**0.8310**	0.7628	0.8298	**0.8038**
vote	—	—	0.9587	0.9564	0.9589	0.9563	**0.9592**	**0.9565**
breast-cancer	0.7107	0.7038	0.7156	0.7023	0.7134	0.7022	**0.7163**	**0.7041**
mushroom	0.9972	**0.9972**	1.0000	0.9948	0.9986	0.9921	**1.0000**	0.9954

4.3 统计检验

为探究 FSRSWOA 算法的泛化性能，实验选择 Friedman 检验与 Nemenyi 后续检验来对实验用到的所有算法进行比较。

在假设开始之前，需要根据表 3 中各算法的分类精度对各算法的性能进行排序，并为其赋予序值 1，2，3，…，若性能相同则平分序值，结果如表 4 所列。

若算法性能相同，则其平均序值也应该相同。然而从表 4 中各算法的平均序值来看，各算法的平均序值显然不同，则根据 Friedman 检验可知，“H_0：所有算法的性能相同”这个假设被拒绝，需要进行后续检验，以得到具体算法之间的差异。实验中采用的是 Nemenyi 后续检验。

首先使用 Nemenyi 检验计算出平均序值差别的临界值 CD：

$$CD=q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{17}$$

其中，k 为算法的个数，N 为数据集的个数。

实验中 $k=4$，当 $\alpha=0.1$ 时，查表得 $q_\alpha=2.2910$，于是得到临界值 $CD=1.3227$。根据表 4 中各算法的平均序值，计算 FSRSWOA 算法与各对比算法之间平均序值的差值。

表 4 不同数据集上的算法性能排序结果

Table 4 Sorting results of performance of algorithms on different datasets

数据集	FARCF		FSARSR		PSORSFS		FSRSWOA	
	Best	Avg	Best	Avg	Best	Avg	Best	Avg
zoo	4	4	1	3	3	2	2	1
wine	4	4	2	3	3	2	1	1
spectf-heart	4	3.5	3	2	2	3.5	1	1
vehicle	4	4	3	3	2	2	1	1
bands	3	1	4	4	2	3	1	3
flare	4	2	2	4	3	3	1	1
tic-tac-toe	4	4	2	3	1	2	3	1
vote	4	4	3	2	2	3	1	1
breast-cancer	4	2	2	3	3	4	1	1
mush-room	4	1	1.5	3	3	4	1.5	2
average value	3.9000	2.9500	2.3500	3.0000	2.4000	2.8500	1.3500	1.3000

实验分别从算法的最佳分类精度和平均分类精度两个方面来比较各算法的性能。

1）就最佳分类精度的平均序值的差值（与 FARCF 算法、FSARSR 算法以及 PSORSFS 算法的差值分别为 2.5500，1.0000 和 1.05）而言，FSRSWOA 算法仅与 FARCF 算法的平均差值大于 CD 值，即 FSRSWOA 算法得到最佳分类精度的能力显著优于 FARCF 算法，而与其他算法在平均序值上的差值并未大于 CD 值，即 FSRSWOA 优于 FSARSR 算法和 PSORSFS 算法，但性能提升还不够明显。

2）就平均分类精度的平均序值的差值（与 FARCF 算法、FSARSR 算法以及 PSORSFS 算法的差值分别为 1.6500，1.7000 和 1.5500）而言，FSRSWOA 算法的平均序值与其他所有算法的差值均大于 CD 值，即 FSRSWOA 算法的性能显著优于其他算法。

上述统计检验结果说明，FSRSWOA 算法在得到最佳分类性能方面有着微弱优势；在平均分类性能方面，其相比其他算法有着明显的优势。总体来说，FSRSWOA 算法在特征选择方面有一定的优势。

结束语 本文通过一种简单有效的手段将原始的鲸鱼优化算法运用到离散空间；引入群体划分策略，在精英群体的收敛过程中加入扰动，以降低陷入局部最优的风险，并给出了改进的鲸鱼算法；进而将改进的鲸鱼算法与粗糙集理论相结合，提出了新的特征选择算法。该算法利用鲸鱼算法收敛速度快、全局寻优能力强的特点，结合粗糙集原理中属性依赖度的概念，能较快速地找到近似最优特征子集。统计检验的结果表明，该方法能够选出有效的特征子集，并使得后续的学习模型有明显的性能提升。适应度函数和鲸鱼的位置更新方式是影响特征选择结果的关键因素，下一步的工作将考虑进一步提高演化算法的效率以适应大数据处理的需求，并可以考虑引入并行化的思想来提升计算速度。

参 考 文 献

[1] ZHANG D，CHEN S，ZHOU Z. Constraint Score：A new filter method for feature selection with pairwise constraints[J]. Pattern Recognition，2008，41(5)：1440-1451.

[2] SOLORIO-FERNANDEZ S，MARTINEZ-TRINIDAD J F，CARRASCO-OCHOA J A. A new unsupervised spectral feature selection method for mixed data：A Filter Approach[J]. Pattern

Recognition, 2017, 72: 314-326.
[3] LI J D, LIU H. Challenges of feature selection for big data analytics[J]. IEEE Intelligent Systems, 2017, 32(2): 9-15.
[4] MIAO J Y, NIU L F. A survey on feature selection [J]. Procedia Computer Science, 2016, 91: 919-926.
[5] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods[J]. Computers and Electrical Engineering, 2014, 40(1): 16-28.
[6] LI M, KAMILI M. Research on feature selection methods and algorithms[J]. Computer Technology and Development, 2013 (12): 16-21.
[7] LEAS S, CANUTO AM D P. Filter-based optimization techniques for selection of feature subsets in ensemble systems[J]. Expert Systems with Applications, 2014, 41(4): 1622-1631.
[8] YANG P, LIU W, ZHOU B B, et al. Ensemble-based wrapper methods for feature selection and class imbalance learning[C]// Advances in Knowledge Discovery and Data Mining. 2013, 7818: 544-555.
[9] HAMED T, DARA R, KREMER S C. An Accurate, fast embedded feature selection for SVMs[C]// Proceedings of the 2015 International Conference on Machine Learning and Applications. Piscataway, NJ: IEEE, 2015: 135-140.
[10] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.
[11] YU Y, PEDRYCZ W, Miao D. Neighborhood rough sets based multi-label classification for automatic image annotation[C]// Proceedings of the 2013 Ifsa World Congress and Nafips Meeting. Piscataway, NJ: IEEE, 2013: 1373-1387.
[12] WANG C, SHAO M, He Q, et al. Feature subset selection based on fuzzy neighborhood rough sets[J]. Knowledge-Based Systems, 2016, 111: 173-179.
[13] ZHOU J, PEDRYCZ W, Miao D. Shadowed sets in the characterization of rough-fuzzy clustering[J]. Pattern Recognition, 2011, 44(8): 1738-1749.
[14] BANERJEE A, MAJI P. Rough sets and stomped normal distribution for simultaneous segmentation and bias field correction in brain MR images[J]. IEEE Transactions on Image Process, 2015, 24(12): 5764-5776.
[15] ALBANESE A, PAL S K, PETROSINO A. Rough sets, kernel set, and spatiotemporal outlier detection[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(1): 194-207.
[16] ZHOU B, CHEN L, JIA X. Information retrieval using rough set approximations[M]// ICTs and the Millennium Development Goals. Springer US, 2014: 185-197.
[17] HU Q H, ZHAO H, YU R D. Efficient symbolic and numerical attribute reduction with neighborhood rough sets[J]. Pattern Recognition and Artificial Intelligence, 2008, 21(6): 730-738.
[18] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[C]// Proceedings of the 1991 Intelligent Decision Support-handbook of Applications and Advances of the Rough Sets theory. Dordrecht: Kluwer Academic Publisher, 1991: 331-362.
[19] VIEGAS F, ROCHA L, GONÇALVES M, et al. A Genetic Programming approach for feature selection in highly dimensional skewed data[J]. Neurocomputing, 2018, 273: 554-569.
[20] OH I S, LEE J S, MOON B R. Hybrid genetic algorithms for feature selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(11): 1424-1437.
[21] MOAYEDIKIA A, ONG K, BOO Y L, et al. Feature selection for high dimensional imbalanced class data using harmony search [J]. Engineering Applications of Artificial Intelligence, 2017, 57: 38-49.
[22] MITIC M, VUKOVIC N, PETROVIC M, et al. Chaotic fruit fly optimization algorithm [J]. Knowledge-Based Systems, 2015, 89(C): 446-458.
[23] CHEN Y M, MIAO D Q, WANG R Z. A rough set approach to feature selection based on ant colony optimization[J]. Pattern Recognition Letters, 2010, 31(3): 226-233.
[24] XUE B, ZHANG M, BROWNE W N. Particle swarm optimization for feature selection in classification: a multi-objective approach[J]. IEEE Transactions on Cybernetics, 2013, 43(6): 1656-1671.
[25] WANG X, YANG J, TENG X, et al. Feature selection based on rough sets and particle swarm optimization[J]. Pattern Recognition Letters, 2007, 28(4): 459-471.
[26] WANG L, QIU T R, HE N, et al. A method for feature selection based on rough sets and ant colonyoptimization algorithm[J]. Journal of Nanjing University(Natural Sciences), 2010, 46(5): 487-493.
[27] CHEN Y, ZHU Q, XU H. Finding rough set reducts with fish swarm algorithm[J]. Knowledge-Based Systems, 2015, 81(C): 22-29.
[28] MIRJALILI S, LEWIS A. The Whale optimization algorithm [J]. Advances in Engineering Software, 2016, 95: 51-67.
[29] WAIKATO M L G. Weka 3: Data Mining Software in Java [EB/OL]. [2018-07-10]. http://www.cs.waikato.ac.nz/ml/weka/.

WANG Sheng-wu, born in 1995, postgraduate, is member of China Computer Federation (CCF). His main research interests include cloud computing and intelligent technology.

CHEN Hong-mei, born in 1971, Ph.D, professor, Ph.D supervisor, is member of China Computer Federation (CCF). Her main research interests include granular calculation, rough sets and intelligent information processing.