文章编号：2095-2163(2021)11-0049-05 中图分类号：TP391 文献标志码：A

对称不确定性和粒子群的高维特征选择算法

林炜星，王宇嘉，陈万芬
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：高维数据中存在着成千上万个特征，大量的特征导致问题搜索空间过大，增加了计算代价，影响了数据分类预测的准确性。为了提高特征选择的效率，本文提出了一种对称不确定性和种群降维机制的粒子群特征选择算法，该算法设计了一种基于对称不确定性指标的初始化方法，降低特征选择的计算代价。通过非支配排序的种群降维机制，减少进化过程中冗余特征的影响。在 5 个公开生物医学的高维数据集上的实验结果表明，该算法能够针对高维数据特征选择问题取得更好的分类精度和更小的最优子集特征个数，并在时间运行方面有一定的优势。

关键词：对称不确定性；非支配排序；降维；高维数据；特征选择

Symmetrical uncertainty and particle swarm algorithm for feture selection on high-dimensional data

LIN Weixing, WANG Yujia, CHEN Wanfen
(School of Electric and Electronic Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 There are thousands of features in high dimensional data, and a large number of features lead to excessive problem search space, increasing the calculation cost, affecting the accuracy of data classification prediction. In order to improve the efficiency of feature selection, a particle group feature selection algorithm for symmetric uncertainty and population reduction mechanism is proposed. The algorithm is designed with an initialization method based on symmetric uncertain index, and reduces the calculation cost of feature selection. The influence of redundant feature during evolution is reduced by non-dominant sorting population reduction mechanisms. Experimental results on 5 public biologically high dimensional data sets show that the algorithm can achieve better classification accuracy and smaller bistel collection characteristics for high dimensional data characteristics, and operate in time aspects have a certain advantage.

【Key words】 symmetric uncertainty; non-dominated sorting; dimensionality reduction; high-dimensional data; feature selection

0 引 言

近年来，特征选择已经成为数据预处理中的一项重要技术。在现实生活中存在多种类型的数据，如不平衡数据、大样本数据及高维数据等。不平衡数据是指数据分类中，各类数据的样本数相差较大。大样本数据往往具有较少的特征，成千上万的样本数量。高维数据则与之相反，其特点是侧重于大量的特征，通常认为拥有 1 000 以上特征的数据集为高维数据集。在高维数据中的无关特征和冗余特征造成计算代价巨大，增加了学习任务的难度，导致许多学习算法的性能表现不佳。因此，针对高维搜索空间，如何降低维数，减小计算代价成为高维特征选择领域的研究重点。

目前，针对分类问题有许多特征选择方法，这些方法可以分为 3 类，即过滤式选择、包裹式选择和嵌入式选择。过滤式选择方法是直接从数据中计算出性能评估指标进行特征选择，通常有较小的计算开销，如 Relief 算法、mRMR 算法等[1-2]；包裹式选择的思想是将学习器的性能作为特征子集的评价准则，因此该方法通常具有最高的分类精度，如 LVW 算法、GeFeS 算法等[3-4]；嵌入式选择方法中，特征选择和学习器训练在同一个优化过程中完成，通常与特定分类任务相关联，不具有通用性[5]。

由于包裹式选择通常能比过滤式选择取得更好的分类结果，近年来将包裹式选择和群智能算法相结合的研究引起了很多关注。文献[6]提出了一种粗糙集与改进鲸鱼优化算法结合的特征选择，通过粗糙集理论去除数据集中冗余信息来保留原始特征的信息，同时将鲸鱼优化算法运用到了离散空间，通过种群划分策略和扰动策略维持种群的多样性；文献[7]提出了一种基于差分进化的灰狼优化算法，

作者简介：林炜星(1995-)，男，硕士研究生，主要研究方向：进化算法；王宇嘉(1979-)，女，博士，副教授，主要研究方向：进化算法和多目标优化；陈万芬(1995-)，女，硕士研究生，主要研究方向：进化算法。
通讯作者：王宇嘉 Email：yjwangamber@ sues.edu.cn

收稿日期：2021-05-23

引入差分进化机制对种群进行变异操作,采用调节缩放因子和交叉概率因子,提高算法搜索性能,该方法已成功应用于高维医学数据的特征选择问题上;文献[8]提出了一种改进的混合蛙跳算法进行特征选择,该算法在传统蛙跳算法的基础上混合了混沌记忆权重因子、绝对平衡组策略及自适应转移因子的改进方法,有效地提高了特征选择的准确率;文献[9]提出了一种基于改进斑点鬣狗算法的同步优化特征选择,利用自适应差分进化算法、混沌初始化和锦标赛选择策略对斑点鬣狗优化算法进行改进,增强其搜索寻优能力与求解精度。

粒子群优化算法(particle swarm optimization, PSO)作为群智能算法也被应用在许多特征选择问题上[10]。Binh 等将粒子群算法应用于特征选择时,提出了一种基于 PSO 的特征选择方法,并有效结合了过滤式选择和包裹式选择的优点,通过设计局部搜索策略和改进适应度函数平衡算法的开发和勘探能力[11];Lane 等人使用统计聚类方法将相似的特征归为一个簇,在粒子寻优中选择每个簇中具有最高概率的特征,实验结果表明该方法可以挑选出较小规模的特征子集,改善了分类准确率[12]。

本文针对高维搜索空间过大带来的过高的计算代价,提出了基于对称不确定性和种群降维机制的粒子群算法(symmetric uncertainty population dimensionality reduction particle swarm optimization, SUPDR-PSO),该算法首先引入了对称不确定性指标去除冗余特征,基于对称不确定性设计了一种改进的种群初始化方法,提升算法的搜索效率;其次,采用非支配排序策略思想,设计了种群降维机制在算法进入停滞时改变子代生成的机制;最后,设计了适用于高维数据的适应度函数,指导算法寻优。实验证明,该算法能有效减小问题搜索空间,提高分类准确率。

1 相关概念介绍

1.1 二进制粒子群算法

粒子群优化算法源于鸟群等生物群体捕食行为的研究,是一种求解连续优化问题的群智能方法[10]。粒子群算法依靠种群内个体之间的协作和信息共享寻找最优解。算法的核心部分包含两个重要的更新公式,即速度更新和位置更新。文献[13]将粒子群算法应用于二进制搜索空间,提出了一种二进制粒子群算法(Binary Particle Swarm Optimization, BPSO)。

对于第 t 代的粒子 $\vec{x}_i^t = [x_{i,1}^t, x_{i,2}^t, \ldots, x_{i,D}^t]^{\mathrm{T}}$($i \in \{1,2,\ldots,NP\}$),BPSO 通过公式(1)更新粒子的速度 $\vec{v}_i^t = [v_{i,1}^t, v_{i,2}^t, \ldots, v_{i,D}^t]^{\mathrm{T}}$。

$$v_{i,j}^{t+1} = \omega * v_{i,j}^t + c_1 * r_1 * (pbest_{i,j}^t - x_{i,j}^t) + c_2 * r_2 * (gbest_j^t - x_{i,j}^t) \tag{1}$$

其中,$j \in \{1,2,\ldots,D\}$;

$pbest_i^t = [pbest_{i,1}^t, pbest_{i,2}^t, \ldots, pbest_{i,D}^t]^{\mathrm{T}}$ 是 $\vec{x}_i^t$ 的历史最优位置;$gbest_i^t = [gbest_1^t, gbest_2^t, \ldots, gbest_D^t]^{\mathrm{T}}$ 是所有粒子在第 t 代的历史最优位置;ω 是惯性权重,平衡种群的搜索能力和探索性能;c_1 和 c_2 是加速因子;r_1 和 r_2 是均匀分布在 [0,1] 间的随机数。

每一代中,粒子速度 $v_{i,j}^{t+1}$ 限制在 $[-v_{max}, v_{max}]$。在 BPSO 中,位置更新通过 *sigmoid* 函数转换实现,式(2)和式(3):

$$s(v_{i,j}^{t+1}) = \frac{1}{1 + e^{-v_{i,j}^{t+1}}} \tag{2}$$

$$x_{i,j}^{t+1} = \begin{cases} 1, \text{if rand} \leqslant s(v_{i,j}^{t+1}) \\ 0, \text{otherwise} \end{cases} \tag{3}$$

其中,*rand* 是均匀分布在[0,1]之间的随机数,$S(\cdot)$ 是 *sigmoid* 函数。

将 BPSO 算法应用于特征选择时,1 表示该特征被选择,0 表示该特征不被选择。

1.2 特征排序

根据数据集中的特征与类标签之间的相关性强弱,可以将特征分为无关特征、弱相关冗余特征、弱相关非冗余特征和强相关特征。特征选择的任务就是选出最优特征子集,即弱相关非冗余特征和强相关特征,在保证特征子集规模尽可能小的前提下,保证良好的准确率。本文采用对称不确定性指标(symmetrical uncertainty, SU)来衡量特征的相关性,其是信息增益的归一化版本,*SU* 的值越大,说明特征 F 与类标签 C 的相关性越大,*SU* 计算公式(4)和公式(5)[14]:

$$SU(F,C) = \frac{IG(F \mid C)}{H(F) + H(C)} \tag{4}$$

$$IG(F \mid C) = H(F) - H(F \mid C) \tag{5}$$

其中,$H(F)$ 是特征 F 的熵;$H(F \mid C)$ 是给定类标签 C 下 F 的条件熵;*SU* 的取值为 [0,1],1 代表最相关特征。

2 基于对称不确定性和粒子群算法的特征选择方法

2.1 基于对称不确定性的初始化方法

基于种群的算法通常采用随机初始化方法,而

随机初始化产生的种群往往集中在目标空间的部分区域，导致种群多样性差，解的质量低下。考虑到高维数据特征选择问题搜索空间巨大，生成的解在不断趋向 Pareto 前沿的优化过程中，难免需要分配更多的计算资源，产生了巨大的计算代价。因此，针对该现象本文提出基于对称不确定性的初始化方法，其具体流程如下：

步骤 1　使用 SU 评估每个特征与类标签之间的相关性，特征与类标签的相关性越大，则该特征就越重要；

步骤 2　根据特征的重要度，从原始特征集中删除无关特征和弱相关特征，设置阈值 θ 来删除无关特征和弱相关特征，设置 $\theta = 0.1 * \mathrm{cdmax}$，其中 cdmax 为当前数据集的最大特征重要度；

步骤 3　将剩余的特征按照特征重要度进行排序，分为两个集合，分别记为 $High - su$（SU 值高）和 $Low - su$（SU 值低）；

步骤 4　生成随机数 L，控制初始个体的特征数目，从 $High - su$ 中选出 $ceil(rand(0.5,1) * L)$ 个特征，从 $Low - su$ 中选出 $L - ceil(rand(0.5,1) * L)$ 个特征，将选取的特征构建成一个个体；

步骤 5　重复步骤 4 直到完成种群的初始化。

2.2　基于非支配排序的种群降维机制

由于算法采用的是二进制编码方式，随着迭代的进行，决策空间中出现多个完全相同的解。随着大量无关特征和冗余特征的影响，极易造成 Pbest 在更新中保持不变，导致算法优化过程出现停滞现象，因此引入非支配排序策略对特征维度进行约简。

基于非支配排序的种群降维机制是根据当前种群个体的特征数以及分类准确率，通过非支配排序理论对种群中的个体进行排序，取第一前沿面的非支配个体，并统计出现在这些个体中的特征，使用这些特征产生子种群，未出现的特征则在后续的进化过程中忽略。种群降维机制的流程如下：

步骤 1　根据分类准确率和子集特征数目，对当前种群中的所有个体进行排序，取第一前沿的非支配个体，记为集合$\{f1\}$；

步骤 2　统计$\{f1\}$中所有个体携带的特征，将其视为重要特征，没有出现的特征在后续进化过程中不再出现；

步骤 3　利用轮盘赌算法生成子代种群。

SUPDR-PSO 算法的流程图如图 1 所示。

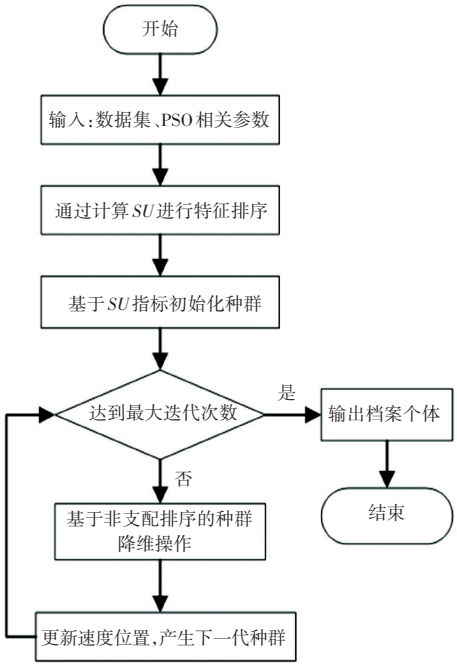

图 1　SUPDR-PSO 算法流程图

Fig. 1　Flow chart of SUPDR-PSO

3　实验结果分析

3.1　测试数据集

为了验证本文所提算法的有效性，使用了公开生物医学数据集 Kent Ridge Biomedical Dataset（http://leo.ugr.es/elvira/DBCRepository/），分别为数据集的特征个数、样例个数及类别数，见表 1。对于每个数据集，样例被划分为两个集合，即 70% 的样例作为训练集，30% 的样例作为测试集，在算法执行过程中，采用十折交叉验证法进行评价。采用 K 近邻算法作为分类器评价分类性能，其中 $K = 5$。每个算法分别独立运行 30 次，分别对比特征数和分类精度的平均值。

表 1　数据集信息

Tab. 1　Information of dataset

数据集	特征数	样例数	类别数
ColonTumor	2 000	62	2
LungCancerOntario	2 880	39	2
DLBCL-Stanford	4 026	47	2
DLBCL-Harvard	6 817	58	2
DLBCL-NIH	7 399	160	2

3.2　参数设置

本文使用 IPSO[15]、IGA[16]、GWODE 和 ISFLA 与

本文算法进行比较。SUPDR-PSO 算法种群大小设置为 200,最大迭代次数设置为 300。所有算法都是在 CPU 为 1.8 GHz Inter Core i7-8550U、内存为 8 G、操作系统为 Windows 10 64 位的 PC 上,在 MATLAB 独立环境下运行 30 次,分别对比特征数和分类精度的平均值。

3.3 结果分析

3.3.1 分类结果分析

分析 4 种群智能算法与本文算法的分类准确率,各算法独立运行 30 次获得的平均分类准确率见表 2,可以看出群智能算法在高维数据特征选择问题上可以成功实现分类,并取得较好的分类准确率。SUPDR-PSO 算法在 ColonTumor、LungCancerOntario、DLBCLHarvard 和 DLBCL-NIH 上均取得了最好的结果,分别为 94.33%、91.77%、86.27% 和 73.00%。在 DLBCLStanford 数据集上,GWODE 取得了最优的分类精度为 99.71%,比本文所提算法高出了 5.12%。因此,本文算法在高维数据特征选择上能取得较好的分类准确率,具有一定的竞争力。

表 2 分类准确率对比

Tab. 2 Comparison of classification accuracy

Datasets	IPSO	IGA	GWODE	ISFLA	SUPDR-PSO
ColonTumor	87.67	86.67	93.42	93.02	**94.33**
LunCancerOntario	70.00	65.50	91.34	75.06	**93.77**
DLBCL-Stanford	78.10	78.80	**99.71**	82.67	94.59
DLBCL-Harvard	71.11	64.33	82.52	73.33	**86.27**
DLBCL-NIH	55.10	56.10	72.53	55.63	**73.00**

本文算法和对比算法取得的特征子集规模的大小,数据为各算法独立运行 30 次获得的平均子集特征数目,见表 3。SUPDR-PSO 算法在 ColonTumor、DLBCL-Stanford、DLBCL-Harvard 和 DLBCL-NIH 上均能取得最小的特征子集规模,证明了种群降维机制的有效性。在 LungCancerOntario 数据集上,GWODE 得到最小的特征子集规模为 4.31,相较于 SUPDR-PSO 算法小了 8.66。特征选择的任务是同时最小化特征子集规模和分类准确率,GWODE 虽取得最小的子集规模,但是由表 2 可以看出,该算法在分类准确率上略小于本文算法,这是由于过分压缩子集规模导致的重要信息丢失从而造成了准确率的下降。

表 3 最优子集特征数对比

Tab. 3 Comparison of feature subset numbers

Datasets	IPSO	IGA	GWODE	ISFLA	SUPDR-PSO
ColonTumor	49.40	38.24	27.10	35.22	**26.33**
LungCancerOntario	56.25	10.22	**4.31**	14.33	12.97
DLBCL-Stanford	49.50	18.43	28.23	15.24	**14.33**
DLBCL-Harvard	51.24	27.62	17.81	27.42	**16.23**
DLBCL-NIH	35.10	32.21	25.33	29.25	**24.55**

3.3.2 运行时间分析

对比算法的平均 CPU 运行时间,各算法独立运行 30 次获得的平均运行时间见表 4,时间单位为 min,可以看出对于高维数据集,各算法的运行时间都比较高,在进行特征选择时,算法的计算代价主要集中在个体的评估上,每次评估都需要建立一个新的模型并计算其准确率,因此计算代价也与特征数目呈现出正比关系。分析表 4 可知,SUPDR-PSO 算法取得了最短的运行时间,其主要原因是本文所提算法在初始化过程中使用的 SU 指标过滤后得到了强相关特征,该过程使种群在趋向 Pareto 前沿时可以分配更少的计算资源,从而提高算法的运行效率。

表 4 运行时间对比

Tab. 4 Comparison of running time

Datasets	IPSO	IGA	GWODE	ISFLA	SUPDR-PSO
ColonTumor	10.91	11.04	8.43	11.56	**5.97**
LungCancerOntario	10.57	10.89	9.87	11.57	**6.12**
DLBCL-Stanford	17.68	18.07	7.59	19.47	**4.64**
DLBCL-Harvard	38.87	38.99	43.23	40.66	**7.87**
DLBCL-NIH	114.68	115.58	13.48	119.68	**10.55**

4　结束语

本文提出了一种基于对称不确定性和粒子群算法的特征选择方法。实验结果表明，针对高维数据的分类问题，本文所提出的方法能够有效地减小分类错误率，提高分类准确率。该方法可以作为一种实用的数据预处理方法来帮助医学领域进行特征选择工作，更好地帮助医学诊断。

参考文献

[1] KIRA K, RENDELL L A. The feature selection problem: traditional methods and a new algorithm[C]// Tenth National Conference on Artificial Intelligence. San Jose, CA : AAAI Press, 1992: 129-134.

[2] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(8): 1226-1238.

[3] LIU H, SETIONO R. Feature Selection and Classification - A Probabilistic Wrapper Approach [C]// Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Proceedings of the Ninth International Conference, Fukuoka, Japan: 1996: 419-424.

[4] SAHEBI G, MOVAHEDI P, EBRAHIMI M, et al. GeFeS: A generalized wrapper feature selection approach for optimizing classification performance [J]. Computers in Biology and Medicine, 2020: 103974.

[5] HAMED T, DARA R, KREMER S C. An Accurate, Fast Embedded Feature Selection for SVMs [C]// International Conference on Machine Learning & Applications. Detroit, MI, USA : IEEE, 2015: 135-140.

[6] 王生武, 陈红梅. 基于粗糙集和改进鲸鱼优化算法的特征选择方法[J]. 计算机科学, 2020, 47(2): 44-50.

[7] 王俊,冯军,张戈,等. 基于改进灰狼优化算法的医学数据特征选择应用研究[J]. 河南大学学报: 自然科学版, 2020(5): 570-578.

[8] 代永强,郭小燕,王敏,等. 基于混合蛙跳算法的高维生物医学数据特征选择方法[J/OL]. 计算机应用研究: 1-8[2021-04-01].https://doi.org/10.19734/j.issn.1001-3695.2020.04.0115.

[9] 贾鹤鸣,姜子超,李瑶,等. 基于改进斑点鬣狗算法的同步优化特征选择[J/OL]. 计算机应用: 1-11[2021-03-31].http://kns.cnki.net/kcms/detail/51.1307.tp.20201209.1510.010.html.

[10] KENNEDY J, EBERHART R. A new optimizaer using particle swarm theory [C]//Proceedings of the sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan: IEEE, 1995: 39-43.

[11] TRAN B, ZHANG M, BING X. A PSO based hybrid feature selection algorithm for high-dimensional classification[C]// 2016 IEEE Congress on Evolutionary Computation (CEC). Vancouver, BC, Canada : IEEE, 2016: 3801-3808.

[12] LANE M C, BING X, LIU I, et al. Gaussian Based Particle Swarm Optimisation and Statistical Clustering for Feature Selection [C]// European Conference on Evolutionary Computation in Combinatorial Optimization. Granada, Spain: Spinger, 2014: 133-144.

[13] KENNEDY J, EBERHART R C. A discrete binary version of the particle swarm algorithm [C]//1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation Orlando, FL, USA: IEEE, 1997: 4104-4108.

[14] YU L, LIU H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. The Journal of Machine Learning Research, 2004, 5(12): 1205-1224.

[15] KARTHIGA R, MANGAI S. Feature Selection Using Multi-Objective Modified Genetic Algorithm in Multimodal Biometric System[J]. Journal of Medical Systems, 2019, 43(7): 214.

[16] GUNASUNDARI, JANAKIRAMAN, MEENAMBAL. Multiswarm heterogeneous binary PSO using win-win approach for improved feature selection in liver and kidney disease diagnosis. [J]. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society, 2018, 12(70): 135-54.

(上接第 48 页)

[8] 尚明姝,王克朝. 一种基于 SURF 和 BRIEF 的图像配准算法[J]. 微电子学与计算机, 2020, 37(10): 59-63.

[9] 木拉提·哈米提,张岁霞,等. KNN 分类器在新疆维吾尔药材图像分类中的应用[J]. 新疆医科大学学报, 2015, 38(7): 799-804.

[10] 吴映铮,杨柳涛. 基于 HOG 和 SVM 的船舶图像分类算法[J]. 上海船舶运输科学研究所学报, 2019, 42(1): 58-64.

[11] 高峥,徐震. 基于多元回归 KNN 的油田缺失数据填充方法[J]. 信息技术, 2020, 44(4): 79-83.

[12] 陈玉林,李戈理,杨智新,等. 基于 KNN 算法识别合水地区长期储层岩性岩相[J]. 测井技术, 2020, 44(2): 182-185.

[13] 冯泽安,王鹏. 基于多分类模型加权投票法的人脸微笑检测[J]. 计算机技术与发展, 2019, 29(2): 81-86.

[14] 薛飞,刘立群. 基于 OTSU 算法的苹果果实病斑图像分割方法[J]. 计算机技术与发展, 2020, 30(12): 181-186.

[15] 李澎林,邹嘉程,李伟. 基于 HOG 和特征描述子的人脸检测与跟踪[J]. 浙江工业大学学报, 2020, 48(2): 133-140.

[16] 孟金龙,丁超洋,周慧,等. 基于 SVM 的图像分类算法研究[J]. 数字技术与应用, 2017(10): 123-124.

[17] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size [J]. arXiv preprint arXiv: 1602.07360, 2016.

[18] CHE M L, CHE M J, CHAO Z H , et al. Traffic Light Recognition for Real Scenes Based on Image Processing and Deep Learning, Computing and Informatics, 2020, 39(3): 439-463.