山西大学学报(自然科学版)47(1):1—8,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023135

基于互信息和遗传算法的特征选择算法

张婧[1],曹峰[2*],董毓莹[2],张超[2],余银中[3],唐超[4]
(1.太原学院 数学系,山西 太原 030032;
2.山西大学 计算机与信息技术学院,山西 太原 030006;
3.安徽华治新能源科技有限公司,安徽 合肥 230601;
4.合肥学院 人工智能与大数据学院,安徽 合肥 230601)

摘 要:本文提出了一种新的基于互信息和遗传算法的监督、封装型特征选择算法。该算法设计了基于互信息的特征之间以及特征与类之间的相关性度量指标,并结合遗传算法具有的较强的全局寻优能力,在候选特征空间中寻找特征间相关性低,特征与类相关性高且分类精度高的全局最优特征子集。本文在10个标准数据集上,与8个基于相关性的特征选择算法进行了对比实验。在3个分类器下,本文算法对应的平均分类精度分别为88.98%,87.5%和86.95%,优于所有对比算法。结果表明,本文算法可以有效降低原始特征集的维数并提升分类器的精度。

关键词:特征选择;相关性;熵;互信息;遗传算法
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)01-0001-08

Feature Selection Algorithm Based on Mutual Information and Genetic Algorithm

ZHANG Jing[1], CAO Feng[2*], DONG Yuying[2], ZHANG Chao[2], YU Yinzhong[3], TANG Chao[4]
(1. *Department of Mathematics, Taiyuan University, Taiyuan* 030032*, China;*
2. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China;*
3. *Anhui Huaye New Energy Technology, Hefei* 230601*, China;*
4. *School of Artificial Intelligence and Big Data Studies, Hefei College, Hefei* 230601*, China*)

Abstract: A novel feature selection algorithm using mutual information and genetic algorithm is presented in this paper. The algorithm designed the metrics for measuring the correlation between features and that between features and classes based on mutual information. By combining the strong global optimization capability of genetic algorithms, it can search for a globally optimal feature subset in the candidate feature space, characterized by low inter-feature correlation, high feature-to-class correlation, and high classification accuracy. In this paper, comparative experiments were conducted on 10 standard datasets using 8 correlation-based feature selection algorithms. Under 3 classifiers, the algorithm proposed in this paper achieves average classification accuracies of 88.98%, 87.5%, and 86.95%, respectively, outperforming all the comparative algorithms. The experimental outcomes demonstrate the effectiveness of the proposed algorithm in significantly reducing the dimensionality of the original feature sets while enhancing the accuracies of classifiers.

Key words: feature selection; correlation; entropy; mutual information; genetic algorithm

收稿日期:2023-06-03;**接受日期**:2023-08-04
基金项目:国家自然科学基金(62072291;62272284);安徽省自然科学基金(2008085MF202)
作者简介:张婧(1982-),女,山西五台人,硕士,副教授,研究方向为数据挖掘。E-mail:zj6amanda@163.com
* **通信作者**:曹峰(CAO Feng),E-mail:caof@sxu.edu.cn
引文格式:张婧,曹峰,董毓莹,等.基于互信息和遗传算法的特征选择算法[J].山西大学学报(自然科学版),2024,47(1):1-8. DOI:10.13451/j.sxu.ns.2023135

0 引言

大数据时代已经到来,各个行业领域都在不断的产生各种类型的海量高维数据,如证券市场交易数据、多媒体图形图像视频数据、生物特征数据等。海量高维数据提供了更加准确可靠的信息,有助于我们实施精准的数据分析、建模以及决策。但是,随着数据维数的增加,诸多问题也随之而来。首先,维数过高会导致“维数灾难”[1-2],即随着数据维数的增加,计算量呈指数级增长;其次,在分类问题中会出现“Huges”现象[3],即随着参与建模的数据维数的增加,分类精度会出现“先增后降”的现象;另外,数据维数过高容易导致学习算法过拟合,从而降低算法的泛化能力[4]。数据降维可以有效的解决以上问题,并已成为数据挖掘和机器学习领域重要的研究方向之一。

数据降维主要包括特征提取和特征选择[5]。特征提取通过数学变换,将高维数据映射到低维子空间中,得到一组新的抽象程度更高的特征。特征选择旨在从原始特征集中选择一个最优特征子集,该子集不仅可以降低数据的维数,而且能够提高学习算法的性能[6]。与特征提取不同,特征选择不产生新的特征,具有较好的可解释性。目前,众多学者已经对特征选择算法进行了深入的研究,将其归纳为一个搜索寻优问题[7]。但是最小特征子集搜索问题是一个NP(Non-deterministic Polynomial)问题,因此,特征选择算法主要通过特征重要性评价函数和搜索策略来寻找近似最优的特征子集[8]。如何构建特征重要性评价函数以及如何优化搜索策略是特征选择研究的关键问题。

相关性度量是特征选择算法用于构建特征重要性评价函数的常用指标。基于相关性度量的特征选择算法主要通过选择与类高度相关的特征,去除与类相关性较低的特征,得到一组近似最优的特征子集。1992年,Kira等提出了Relief算法,基于特征对近距离样本的区分能力度量特征和类的相关性,根据特征和类的相关性赋予特征不同的权重,权重小于某个阈值的特征将被移除[9]。该算法比较简单,运行效率较高,但是仅适用于二分类问题。因此,Kononenko等对Relief算法进行了改进,提出了可以解决多分类问题的ReliefF算法。当存在大量无关特征或错误标记时,该算法的性能会受到很大的影响[10-11]。Yu等提出了一种基于对称不确定性的快速过滤特征选择算法,该算法首先通过特征和类之间的对称不确定性值选择相关特征,然后在相关特征中剔除冗余特征[12]。Fleuret将条件互信息引入到特征选择算法中,提出基于条件互信息最大化准则的特征选择算法,该算法在选择特征时既考虑特征与类别的相关性,同时也利用条件互信息对冗余部分进行了最大化处理,通过条件互信息最大化准则(CMIM)判断特征之间相对类的信息冗余度的差异,逐步选择信息量大且冗余度低的特征[13]。Hoque等提出一种基于互信息的特征选择算法(MIFS),该算法结合特征与特征之间的互信息以及特征与类之间的互信息寻找特征的最优特征子集,通过最大化特征之间的相关性减少冗余的特征[14]。

以上特征选择算法,主要通过启发式搜索策略寻找全局最优特征子集。启发式搜索策略虽然效率较高,但是很容易陷入局部最优解。为了更好的获得全局最优解,启发式随机搜索策略被广泛应用于特征选择算法中,其中遗传算法最具代表性。Dong等提出了一种融合遗传算法与粒计算的特征选择算法,该算法结合改进的邻域粗糙集和遗传算法优化最优特征子集,进而获得较高的分类精度[15]。裴作飞等提出了CS-GA(Chi Square-Genetic Algorithm)混合特征选择算法,该算法首先采用卡方检验选择相关性高的特征,再通过自适应遗传算法搜索最优特征子集[16]。Xue等提出了一种基于多目标二进制遗传算法的特征选择算法,该算法采用自适应策略优化遗传算法的选择算子,实验结果表明,该特征选择算法在大规模数据集上体现出良好的性能[17]。以上特征选择算法并没有考虑遗传算法迭代过程中所选特征与特征以及特征与类之间的相关性,可能导致所选特征集仍具有较高的冗余度。

针对以上不足之处,本文提出了一种基于互信息相关性度量和遗传算法的监督封装型特征选择算法,该算法基于互信息构建特征之间以及特征与类之间的相关性度量指标,并用于

构建遗传算法适应度函数。同时,本文算法将分类器的分类精度作为遗传算法适应度函数构建的一部分,通过遗传算法在特征空间中搜索特征间相关性低,特征与类相关性高且分类精度高的全局最优特征子集。

1 基础理论

1.1 遗传算法

遗传算法是在二十世纪六七十年代由美国密歇根大学的Holland教授创立。六十年代初,Holland在设计人工自适应系统时提出应借鉴遗传学基本原理模拟生物自然进化的方法。1975年,Holland出版了第一本系统阐述遗传算法基本理论和方法的专著,提出了遗传算法理论研究和发展中最重要的模式理论[18]。同年,De Jong完成了大量基于遗传算法思想的纯数值函数优化计算实验的博士论文,为遗传算法及其应用打下了坚实的基础[19]。1989年,Goldberg对遗传算法做了全面系统的总结和论述,奠定了现代遗传算法的基础[20]。

遗传算法(Genetic Algorithm,GA)是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型,是一种通过模拟自然进化过程搜索最优解的方法[21],具体步骤如下:

步骤1:编码染色体,随机产生种群;

步骤2:根据适应度函数计算种群中每条染色体的适应度值;

步骤3:判断是否满足停止准则,一般选择最大迭代次数作为停止准则;

步骤4:依据某种策略选择适应度值高的染色体直接进入下一代种群;

步骤5:依据一定的交叉概率选择成对染色体进行交叉操作。每两个染色体交换部分染色体后产生子代染色体,代替父代染色体进入新的种群;

步骤6:依据一定的概率选择染色体进行变异操作,发生变异的染色体代替原来的染色体进入新的种群。

步骤7:更新种群,选择、交叉和变异操作之后,产生新的种群,返回步骤2。

1.2 互信息

1948年Shannon建立了信息论理论[22]。根据信息论,熵,又名信息熵,用于描述随机事件的不确定程度。随机事件的不确定度越大,信息熵越大;反之,信息熵越小。如果X是一个用于描述随机事件的随机变量,其概率分布为:

$$P(X=x)=p(x),x\in X。\tag{1}$$

随机变量X的熵定义为:

$$H(X)=-\sum_{x\in X}p(x)\log p(x)。\tag{2}$$

联合熵用来表示两个随机变量X,Y的联合分布,用$H(X,Y)$表示:

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x,y)。\tag{3}$$

条件熵用来表示在随机变量Y已知的情况下,随机变量X的不确定性,公式如下:

$$H(X|Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x|y)。\tag{4}$$

信息熵、联合熵以及条件熵的关系如下:

$$H(X|Y)=H(X,Y)-H(Y)。\tag{5}$$

互信息是衡量随机变量X,Y之间相互依赖程度的度量,即在已知Y的情况下,X的不确定性减少的程度,其公式如下:

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}。\tag{6}$$

互信息可以看成是一个随机变量中包含的关于另一个随机变量的信息量,或者说是一个随机变量由于已知另一个随机变量而减少的不确定性,简而言之,就是两个事件集合之间的相关性。如果两个随机变量的互信息的值比较大,则这两个随机变量的相关性就比较大,互信息和信息熵的关系如下:

$$\begin{aligned}&I(X;Y)=\\&H(X)-H(X|Y)=\\&H(Y)-H(Y|X)=\\&H(X)+H(Y)-H(X,Y)。\end{aligned}\tag{7}$$

互信息可以有效的度量随机变量相关性,因此被广泛地应用于机器学习任务中特征之间相关性的度量。众多学者将互信息引入特征选择中,提出了一系列基于互信息的特征选择算法,验证了互信息在度量特征之间相关性的有效性。基于此,本文基于互信息设计新的特征选择算法。

2 本文算法

基于相关性的特征选择旨在从原始特征集合中找到与类相关性最大,同时特征之间相关性最小的一组特征子集。因此,如何合理地度量相关性是基于相关性的特征选择算法优劣的关键。本文首先基于互信息定义特征集以及特征集与类的相关性度量指标,接着通过遗传算法具有的良好的全局寻优能力,在原始特征集合中寻找最优特征子集。

2.1 基于互信息的相关性度量

定义1 设S为包含n个特征的特征集,A_i和$A_j(i\neq j)$是S中的任意两个特征,特征集S的相关性为:

$$FF(S)=\frac{2}{n(n-1)}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}I(A_i,A_j)=\frac{2}{n(n-1)}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\left[H(A_i)+H(A_j)-H(A_i,A_j)\right]。\tag{8}$$

$FF(S)$通过计算特征集S中的特征之间的互信息的均值,度量特征之间的平均相关程度,$FF(S)$的值越大表示特征集S的相关性越强,特征集S的冗余度越大。

定义2 设S为包含n个特征的特征集,A_i为S中的任意一个特征,C为类标记,特征集S的类相关性为:

$$FL(S,C)=\frac{1}{n}\sum_{i=0}^{n}I(A_i,C)=\frac{1}{n}\sum_{i=0}^{n}\left[H(A_i)+H(C)-H(A_i,C)\right]。\tag{9}$$

$FL(S,C)$通过计算特征集S中的特征与类的互信息的均值,度量特征与类的平均相关程度,$FL(S,C)$的值越大表示特征集S与类的相关性越大。

2.2 算法流程

基于互信息定义了新的相关性度量指标后,利用遗传算法对特征选择问题进行建模,首先采用二进制的形式对遗传算法的染色体进行编码。染色体的长度为候选特征的数目,“1”表示该特征被选择,“0”表示该特征未被选择。

利用遗传算法进行特征选择时,适应度函数是确保遗传算法收敛到最优解的关键,因此,本文结合新的相关性度量指标构建了新的适应度函数。

定义3 设S为染色体x所对应的特征集,C为类标记,acc为基于特征集S的分类精度,α为权重系数,染色体x的适应度函数为:

$$Fitness(x)=\frac{1}{\alpha\times e^{-acc}+(1-\alpha)\times e^{-(FL(S,C)-FF(S))}}(x)。\tag{10}$$

新的适应度函数以适应度值最大作为最优染色体的评价标准。由公式(10)可以看出,适应度值大要求染色体对应的特征集具有较大的类相关性、较小的特征集相关性以及较大的分类精度。当种群收敛时,最优的染色体对应的特征集就是最优特征子集。本文提出的特征选择算法的算法描述如下:

算法1 基于互信息和遗传算法的特征选择算法

输入:带有类标记的数据集D。

输出:适应度值最大的染色体所对应的特征集。

1. 从数据集D中获取特征集S和类标记L;

2. 初始化种群中染色体的个数*IndNum*,迭代次数*GenNum*,交叉概率P_c,变异概率P_m;

3. 随机产生*IndNum*个长度等于总特征数的二进制染色体组成第一代种群;

4. while *GenNum*<Max*GenNum*

5. for $i=1,2,\cdots,$ *IndNum*

6. 根据公式(10)计算染色体的适应度值;

7. end for

8. 采用锦标赛选择策略进行最优个体选择操作;

9. 以一定的概率P_c选择单个二进制位进行交叉操作;

10. 以一定的概率P_m进行0、1互换变异操作;

11. 产生新的种群;

12. end while

2.3 算法复杂度分析

本文算法的时间复杂度主要取决于染色体适应度值的计算。适应度值的计算过程主要分为以下3个步骤:(1)利用公式(8)计算特征集

中任意两个特征之间的互信息;(2)利用公式(9)计算特征集中任意一个特征与类之间的互信息;(3)计算基于特征集的分类精度。在3个步骤中,时间复杂度最高的是计算特征集的相关性和基于特征集的分类精度。特征集的相关性的时间复杂度为$O(m^2n^2)$,其中m表示特征维数。因此,本文算法的时间复杂度与使用的分类器紧密相关。例如,当使用支持向量机分类器时,支持向量机的时间复杂度最高为$O(mn^2)$,此时本文算法的时间复杂度最高为$O(m^2n^2+mn^2)$。

3 实验结果

3.1 实验数据及参数设置

为了验证本文所提特征选择算法的可行性及性能,使用UCI公共数据集进行了对比实验。表1对实验数据进行了具体描述。实验对比算法包括8个基于相关性的特征选择算法:mRMR[23],ReliefF[10],Fisher Score[24],ICAP[25],CIFE[26],JMI[27],MRI[28],WCFR[29]。

表1 10个广泛使用的UCI公共数据集的数据特征描述

Table 1 Feature description of 10 widely used UCI public datasets

Dataset	Attributes	Instances	Classes
Iris	4	150	3
Diabetes	8	768	2
Breast_cancer	9	682	2
Wine	13	177	3
Zoo	16	101	7
Waveform	21	5 000	3
WDBC	30	568	2
WPBC	33	197	2
Ionosphere	34	350	2
Sonar	60	208	2

为了减少随机性导致的实验误差,实验选择了3个分类器对本文算法得到的最优特征集的优劣进行评价,分别为支持向量机(SVM)、K-近邻(KNN,k=5)和决策树(CART)。实验所采用的3个分类器是机器学习领域最优经典的分类器,常用来评价特征选择算法的有效性。SVM是一种快速可靠的线性分类器,在非线性和高维模式识别中表现出强大的优势。KNN是分类器中最简单的分类器之一,具有非常强的普适性。CART采用树结构进行分类,可以有效的提取决策规则,易于理解和解释。实验采用5倍交叉验证的分类性能评估方式,同时为了保证评价的客观性,对比算法计算得到的最优特征集中的特征数量与本文算法相同。遗传算法的参数选择在寻优过程中发挥着重要的作用,本文结合当前遗传算法相关研究成果中的参数设置方案,并结合本文所采用的实验数据,对遗传算法的参数设置为:种群大小为100,交叉概率为0.25,突变概率为0.3,迭代次数为100代。

3.2 实验结果分析

由于本文所提算法是封装式的,最优特征集中特征的数目与分类器的选择紧密相关,因此,本文首先对比了3种不同分类器下最优特征集中特征数目的差异。图1显示了不同数据集下本文算法封装3种分类器时的最优特征集与原始特征集的特征数目的对比情况。

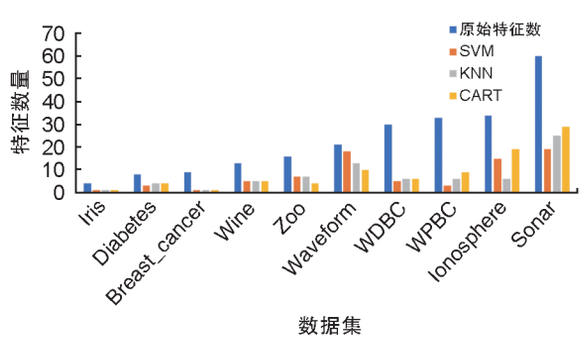

图1 本文特征选择算法采用SVM、KNN和CART分类器的分类精度构建遗传算法适宜度函数时,不同数据集的最优特征数与原始特征数对比

Fig. 1 The feature selection algorithm in this paper uses the classification accuracy of SVM, KNN, and CART classifiers to construct a fitness function for the genetic algorithm. The figure compares the optimal number of features and the original number of features for different datasets

本文特征选择算法与其他8种对比算法在SVM、KNN和CART 3种分类器下的分类精度对比情况如表2—表4所示。图2直观地对比了本文特征选择算法与对比算法在多个数据集下的平均分类精度。

分析实验结果,由图1可以看出,本文特征选择算法所选最优特征集可以有效地压缩原始特征集的特征数目,尤其是高维特征数据。在大多数数据集上,特征数目的减少都在60%以

表2 本文特征选择算法与对比算法在SVM分类器下的分类精度对比(%)

Table 2 Comparison among the classification accuracies of the proposed feature selection algorithm and the comparative algorithms under the SVM classifier (%)

Dataset	Proposed	mRMR	reliefF	Fisher	ICAP	CIFE	JMI	MRI	WCFR
Iris	**96.00**	95.33	95.33	95.33	94.67	94.67	94.67	**96.00**	95.33
Diabetes	**78.43**	66.10	74.19	76.28	75.10	74.97	75.10	75.00	64.89
Breast_cancer	**96.34**	91.65	90.62	90.62	90.91	91.65	91.65	90.16	89.99
Wine	**90.41**	54.32	66.71	66.73	69.02	54.87	88.16	89.95	68.10
Zoo	**97.00**	86.00	87.00	84.00	82.00	82.00	84.00	86.05	85.09
Waveform	**86.99**	85.38	86.92	86.74	86.92	85.14	86.74	86.76	85.08
WDBC	**93.84**	91.56	92.09	91.38	89.98	88.40	92.44	92.10	91.92
WPBC	76.15	76.15	76.15	76.15	75.64	75.64	76.15	76.15	**76.29**
Ionosphere	**96.86**	92.29	95.71	92.86	92.86	90.29	93.71	90.58	95.15
Sonar	**77.78**	61.38	63.72	67.15	56.54	71.09	54.16	53.30	64.15

表3 本文特征选择算法与对比算法在KNN分类器下的分类精度对比(%)

Table 3 Comparison among the classification accuracies of the proposed feature selection algorithm and the comparative algorithms under the KNN classifier (%)

Dataset	Proposed	mRMR	reliefF	Fisher	ICAP	CIFE	JMI	MRI	WCFR
Iris	**96.00**	95.33	95.33	95.33	94.67	95.33	94.67	96.00	95.33
Diabetes	**76.62**	63.49	74.58	75.88	74.71	74.32	75.76	74.10	63.79
Breast_cancer	**96.93**	93.85	90.19	90.19	93.11	93.85	93.85	90.51	90.88
Wine	**92.71**	72.95	68.37	68.89	82.52	75.76	89.87	89.40	70.85
Zoo	**94.00**	88.00	79.00	86.00	84.00	79.00	86.00	88.10	88.10
Waveform	**81.96**	75.43	81.68	81.68	81.68	75.33	81.68	81.12	75.78
WDBC	92.61	92.43	92.61	90.32	89.44	90.51	91.03	**93.50**	89.81
WPBC	**77.21**	71.00	69.01	70.04	64.36	67.41	71.00	69.15	71.61
Ionosphere	**92.57**	89.14	88.57	86.00	84.00	87.71	82.57	83.48	88.90
Sonar	**74.39**	63.26	61.27	64.24	52.57	65.32	54.52	52.78	59.14

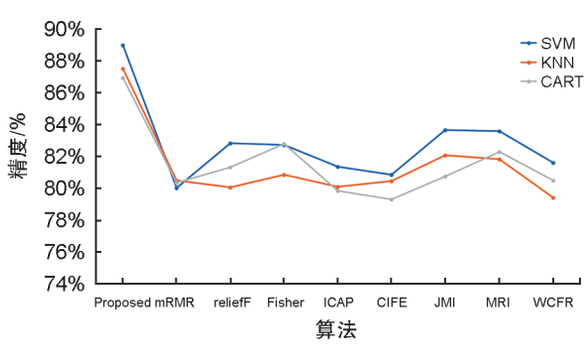

图2 本文特征选择算法与对比算法在SVM、KNN和CART分类器下的平均分类精度对比图

Fig. 2 Comparison among the average classification accuracies of the proposed feature selection algorithm and the comparative algorithms under SVM, KNN, and CART classifiers

上。另外,对比3种封装的分类器,可以发现分类器的封装在特征数目的减少方面无显著的差异,在不同数据集上表现的特征压缩率具有一定的随机性。因此可以认为,本文算法封装的3种分类器在降低特征维数方面都是有效的且无明显差异。

分析表2—表4可以看出,在SVM和KNN分类器下,本文算法与对比算法相比,在9个数据集上的分类精度达到最优。在CART分类器下,本文算法在10个数据集上全部达到最优。图2对比了本文算法与对比算法在多个数据集下的平均分类精度,可以看出,本文算法在3种分类器下的平均分类精度都明显优于对比算法,尤其是在SVM分类器下,本文算法的平均分类精度比最差算法高出约8.97%。本文算法性能较优的主要原因在于:(1)本文算法在最优特征选择的过程中对分类器进行了封装,已经考虑了最优特征对精度的影响;(2)本文算法基于互信息定义的新的相关性度量,可以有效度量特征之间以及特征与类之间的相关性,使得所选最优特征集具有较低的冗余度和较高的类依赖度。

表4 本文特征选择算法与对比算法在CART分类器下的分类精度对比(%)

Table 4 Comparison among the classification accuracies of the proposed feature selection algorithm and the comparative algorithms under the CART classifier (%)

Dataset	Proposed	mRMR	reliefF	Fisher	ICAP	CIFE	JMI	MRI	WCFR
Iris	**72.36**	62.18	66.11	69.10	70.80	69.24	71.18	69.00	62.77
Diabetes	**93.55**	91.65	90.62	90.62	90.91	91.65	91.65	85.59	84.54
Breast_cancer	**95.52**	87.51	89.29	91.59	81.40	82.54	87.60	92.22	93.33
Wine	**96.00**	**96.00**	90.00	95.00	92.00	94.00	95.00	**96.00**	81.00
Zoo	**75.37**	61.91	73.99	71.67	73.27	61.17	74.41	74.78	63.40
Waveform	**94.02**	92.26	92.43	93.13	91.38	89.09	90.32	93.90	92.10
WDBC	**74.65**	67.96	69.40	66.41	64.46	64.28	63.87	65.62	72.68
WPBC	**94.86**	87.71	92.00	86.86	86.00	89.14	83.14	83.75	89.24
Ionosphere	**77.79**	64.22	57.48	71.60	55.55	60.88	58.99	66.89	70.61
Sonar	**95.33**	92.00	92.00	92.00	92.67	91.31	91.31	**95.33**	**95.33**

4 结论

为了通过特征选择进行有效的高维数据降维,进而提高数据分类的精度,本文提出了一种基于互信息和遗传算法的特征选择算法。该算法首先定义了新的特征与特征以及特征与类的相关性度量,并基于定义的相关性度量构建遗传算法适应度评价函数,进而通过遗传算法在原特征集中寻找最优特征子集,确保所选特征子集具有较低的特征冗余度和较高的类依赖度。我们选择了10个UCI数据集,并与8个基于相关性的特征选择算法进行了对比实验。实验结果表明,本文算法在3种分类器下的分类精度都优于对比算法。

在未来的研究工作中,我们将重点关注以下两个方面:(1)利用多变量互信息定义相关性度量,进一步提升算法相关性度量的全面性;(2)将相关性度量与群体智能优化算法中的最新研究成果相结合,进一步改进算法在特征子集寻优方面的性能。

参考文献:

[1] 孙林, 徐枫, 王振, 等. 基于标记权重和mRMR的多标记特征选择[J]. 山西大学学报(自然科学版), 2023, **46**(1): 40−52. DOI: 10.13451/j.sxu.ns.2022105.
SUN L, XU F, WANG Z, *et al*. Multilabel Feature Selection Using Label Weight and MRMR[J]. *J Shanxi Univ Nat Sci Ed*, 2023, **46**(1): 40−52. DOI: 10.13451/j. sxu. ns.2022105.

[2] AYESHA S, HANIF M K, TALIB R. Overview and Comparative Study of Dimensionality Reduction Techniques for High Dimensional Data[J]. *Inf Fusion*, 2020, **59**: 44−58. DOI: 10.1016/j.inffus.2020.01.005.

[3] WANG J X, YE M C, XIONG F C, *et al*. Cross-scene Hyperspectral Feature Selection Hybrid Whale Optimization Algorithm with Simulated Annealing[J]. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2021, **14**: 2473−2483. DOI: 10.1109/JSTARS.2021.3056593.

[4] 殷艳坤. 基于快速密度聚类的特征选择算法[D]. 天津: 南开大学, 2016.
YIN Y K. Feature selection algorithm Based on fast density clustering[D]. Tianjin: Nankai University, 2016.

[5] LI M M, WANG H F, YANG L F, *et al*. Fast Hybrid Dimensionality Reduction Method for Classification Based on Feature Selection and Grouped Feature Extraction[J]. *Expert Syst Appl*, 2020, **150**: 113277. DOI: 10.1016/j.eswa.2020.113277.

[6] HAMMOURI A I, MAFARJA M, AL-BETAR M A, *et al*. An Improved Dragonfly Algorithm for Feature Selection[J]. *Knowl Based Syst*, 2020, **203**: 106131. DOI: 10.1016/j.knosys.2020.106131.

[7] MA J, HAO Z Y, SUN W J. Enhancing Sparrow Search Algorithm via Multi-strategies for Continuous Optimization Problems[J]. *Inf Process Manag*, 2022, **59**(2): 102854. DOI: 10.1016/j.ipm.2021.102854.

[8] KOHAVI R, JOHN G H. Wrappers for Feature Subset Selection[J]. *Artif Intell*, 1997, **97**(1/2): 273−324. DOI: 10.1016/s0004-3702(97)00043-x.

[9] KIRA K, REBDELL L A. Feature Selection Problem: Traditional Methods and a New Algorithm[C]// Proceedings Tenth National Conference on Artificial Intelligence, San Jose, CA: MIT Press, 1992:129−134.

[10] KONONENKO I. Estimating Attributes: Analysis and Extensions of RELIEF[M]//Machine Learning: ECML-94. Berlin, Heidelberg: Springer Berlin Heidelberg,

1994: 171−182. DOI: 10.1007/3-540-57868-4_57.

[11] 丁思凡, 王锋, 魏巍. 一种基于标签相关度的Relief特征选择算法[J]. 计算机科学, 2021, **48**(4): 91−96. DOI: 10.11896/jsjkx.200800025.
DING S F, WANG F, WEI W. Relief Feature Selection Algorithm Based on Label Correlation[J]. *Comput Sci*, 2021, **48**(4): 91−96. DOI: 10.11896/jsjkx.200800025.

[12] YU L, LIU H. Feature Selection for High-dimensional Data: A Fast Correlation-based Filter Solution[C]//Proceedings of the Twentieth International Conference on International Conference on Machine Learning. New York: ACM, 2003: 856−863.

[13] FLEURET F. Fast Binary Feature Selection with Conditional Mutual Information[J]. *J Mach Learn Res*, 2004, **5**: 1531−1555. DOI: 10.1023/b:jiis.0000047395.18103.28.

[14] HOQUE N, BHATTACHARYYA D K, KALITA J K. MIFS-ND: A Mutual Information-based Feature Selection Method[J]. *Expert Syst Appl*, 2014, **41**(14): 6371−6385. DOI: 10.1016/j.eswa.2014.04.019.

[15] DONG H B, LI T, DING R, *et al*. A Novel Hybrid Genetic Algorithm with Granular Information for Feature Selection and Optimization[J]. *Appl Soft Comput*, 2018, **65**: 33−46. DOI: 10.1016/j.asoc.2017.12.048.

[16] 裴作飞, 李兆玉, 王云锋, 等. 基于自适应遗传算法的混合特征选择方法[J]. 计算机应用与软件, 2020, **37**(8): 256−259. DOI: 10.3969/j.issn.1000-386x.2020.08.044.
PEI Z F, LI Z Y, WANG Y F, *et al*. Hybrid Feature Selection Method Based on Adaptive Genetic Algorithm [J]. *Comput Appl Softw*, 2020, **37**(8): 256−259. DOI: 10.3969/j.issn.1000-386x.2020.08.044.

[17] XUE Y, ZHU H K, LIANG J Y, *et al*. Adaptive Crossover Operator Based Multi-objective Binary Genetic Algorithm for Feature Selection in Classification[J]. *Knowl Based Syst*, 2021, **227**: 107218. DOI: 10.1016/j.knosys.2021.107218.

[18] HOLLAND J H. Adaptation in Natural and Artificial Systems[M]. Ann Arbor: University of Michigan Press, 1975: 20−38.

[19] VAFAIE H, DE JONG K A. Genetic Algorithms as a Tool for Feature Selection in Machine Learning[C]//Proceedings of the 4th IEEE International Conference on Tools with AI, Washington DC,1992, 200−203.

[20] GOLDBERG D E. Simple Genetic Algorithms and the Minimal, Deceptive Problem[M]. Los Altos, CA: Morgan Kaufmann, 1987: 74−88.

[21] MOSLEHI F, HAERI A. A Novel Hybrid Wrapper-filter Approach Based on Genetic Algorithm, Particle Swarm Optimization for Feature Subset Selection[J]. *J Ambient Intell Human Comput*, 2020, **11**(3): 1105−1127. DOI: 10.1007/s12652-019-01364-5.

[22] SHANNON C E. A Mathematical Theory of Communication[J]. *Bell Syst Tech J*, 1948, **27**(3): 379−423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[23] PENG H C, LONG F H, DING C. Feature Selection Based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy[J]. *IEEE Trans Pattern Anal Mach Intell*, 2005, **27**(8): 1226−1238. DOI: 10.1109/TPAMI.2005.159.

[24] SUN L, WANG T X, DING W P, *et al*. Feature Selection Using Fisher Score and Multilabel Neighborhood Rough Sets for Multilabel Classification[J]. *Inf Sci*, 2021, **578**: 887−912. DOI: 10.1016/j.ins.2021.08.032.

[25] JAKULIN A, BRATKO I, SMRKE D, *et al*. Attribute Interactions in Medical Data Analysis[M]//Artificial Intelligence in Medicine. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003: 229−238. DOI: 10.1007/978-3-540-39907-0_32.

[26] LIN D H, TANG X O. Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion[M]//Computer Vision-ECCV 2006. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 68−82. DOI: 10.1007/11744023_6.

[27] BENNASAR M, HICKS Y, SETCHI R. Feature Selection Using Joint Mutual Information Maximisation[J]. *Expert Syst Appl Int J*, 2015, **42**(22): 8520−8532. DOI: 10.1016/j.eswa.2015.07.007.

[28] WANG J, WEI J M, YANG Z L, *et al*. Feature Selection by Maximizing Independent Classification Information [J]. *IEEE Trans Knowl Data Eng*, 2017, **29**(4): 828−841. DOI: 10.1109/TKDE.2017.2650906.

[29] ZHOU H F, ZHANG Y, ZHANG Y J, *et al*. Feature Selection Based on Conditional Mutual Information: Minimum Conditional Relevance and Minimum Conditional Redundancy[J]. *Appl Intell*, 2019, **49**(3): 883−896. DOI: 10.1007/s10489-018-1305-0.