



智能系统学报

CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一种新的最大相关最小冗余特征选择算法

李顺勇, 王改变

引用本文:

李顺勇, 王改变. 一种新的最大相关最小冗余特征选择算法[J]. 智能系统学报, 2021, 16(4): 649–661.

LI Shun Yong, WANG Gaibian. New MRMR feature selection algorithm[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 649–661.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202009016>

您可能感兴趣的其他文章

代价敏感数据的多标记特征选择算法

Multi-label feature selection algorithm for cost-sensitive data

智能系统学报. 2019, 14(5): 929–938 <https://dx.doi.org/10.11992/tis.201807027>

基于测度学习支持向量机的钢琴乐谱难度等级识别

Recognition of difficulty level of piano score based on metric learning support vector machine

智能系统学报. 2018, 13(2): 196–201 <https://dx.doi.org/10.11992/tis.201612012>

可拓支持向量分类机

Extension support vector classification machine

智能系统学报. 2018, 13(1): 147–151 <https://dx.doi.org/10.11992/tis.201610019>

中医临床不平衡数据疾病分类方法研究

Research on classification of diseases of clinical imbalanced data in traditional Chinese medicine

智能系统学报. 2017, 12(6): 848–856 <https://dx.doi.org/10.11992/tis.201706046>

基于特征相关的谱特征选择算法

Spectral feature selection based on feature correlation

智能系统学报. 2017, 12(4): 519–525 <https://dx.doi.org/10.11992/tis.201609008>

基于粗糙集相对分类信息熵和粒子群优化的特征选择方法

A feature selection approach based on rough set relative classification information entropy and particle swarm optimization

智能系统学报. 2017, 12(3): 397–404 <https://dx.doi.org/10.11992/tis.201705004>

微信公众平台



关注微信公众号, 获取更多资讯信息

DOI: 10.11992/tis.202009016

网络出版地址: <https://kns.cnki.net/kcms/detail/23.1538.TP.20210406.0853.002.html>

一种新的最大相关最小冗余特征选择算法

李顺勇, 王改变

(山西大学 数学科学学院, 山西 太原 030006)

摘要:传统的基于特征选择的分类算法中, 由于其采用的冗余度和相关度评价标准单一, 从而使得此类算法应用范围受限。针对这个问题, 本文提出一种新的最大相关最小冗余特征选择算法, 该算法在度量特征之间冗余度的评价准则中引入了两种不同的评价准则; 在度量特征与类别之间的相关度中引入了 4 种不同的评价准则, 衍生出 8 种不同的特征选择算法, 从而使得该算法应用范围增大。此外, 由于传统的最大相关最小冗余特征选择算法不能根据用户实际需求的数据维度进行特征选择。所以, 引入了指示向量 λ 来刻画用户实际的数据维度需求, 提出了一种新的目标函数来求解最优特征子集, 利用支持向量机对 4 个 UCI 数据集的特征子集进行了实验, 最后, 利用分类正确率、成对单边 T 检验充分验证了该算法的有效性。

关键词: 特征选择; 冗余度; 相关度; 降维; 分类; 分类正确率; 支持向量机; T 检验

中图分类号: TP181 **文献标志码:** A **文章编号:** 1673-4785(2021)04-0649-13

中文引用格式: 李顺勇, 王改变. 一种新的最大相关最小冗余特征选择算法 [J]. 智能系统学报, 2021, 16(4): 649-661.

英文引用格式: LI Shunyong, WANG Gaibian. New MRMR feature selection algorithm[J]. CAAI transactions on intelligent systems, 2021, 16(4): 649-661.

New MRMR feature selection algorithm

LI Shunyong, WANG Gaibian

(School of Mathematical Sciences, Shanxi University, Taiyuan 030006, China)

Abstract: The application scopes of traditional classification algorithms based on feature selection are limited due to the single evaluation criteria of redundancy and relevance adopted. To solve this problem, this paper proposes a new maximum relevance, minimum redundancy (MRMR) feature selection algorithm, which enlarges its application scope by introducing two different evaluation criteria to measure the redundancy between features of measurement, measuring the correlation between features and categories, and deriving eight different feature selection algorithms. In addition, because the traditional MRMR feature selection algorithms cannot realize feature selection according to the data dimension of users' actual demand, the study also applies an indicator vector λ to achieve that, proposes a new objective function to obtain the optimal feature subset, and conducts experiments on four feature subsets of UCI using a support vector machine. Finally, the study verifies the effectiveness of the algorithm using classification accuracy and pairs of unilateral T-tests.

Keywords: feature selection; redundancy; relevance; dimension reduction; classification; classification accuracy; support vector machines; T-test

特征选择是数据挖掘、机器学习和模式识别

收稿日期: 2020-09-11. 网络出版日期: 2021-04-06.

基金项目: 山西省留学人员科技活动择优资助项目 (2019-13); 山西省基础研究计划项目 (201901D111320); 太原市科技计划研发项目 (2018140105000084); 山西省高等学校精品共享课程项目 (K2020022).

通信作者: 李顺勇. E-mail: lis75@sxu.edu.cn.

中的一项重要技术, 是当前信息领域的研究热点之一^[1-3]。它在数据分析和预处理过程中起着非常重要的作用。特征选择在不改变特征原始表达的基础上, 仅从特征集中筛选最能代表数据特点的最优特征子集。因此, 不仅可以去除不相关和

冗余信息,降低训练样本的维度和分类样本的复杂度,而且能很好地保持原始特征包含的信息,对于人们理解和判断观测来说更加容易。特征选择根据其是否与后续学习算法独立可以分为过滤式和封装式两种。过滤式特征选择方法独立于后续的学习算法,通过数据的本质属性对所有特征进行评分,在此评价过程中不会借用分类模型来完成^[4-5]。其中具有代表性的方法有T检验(T-test)^[6]、Fisher score^[7]、信息增益(information gain, IG)^[8]等。但是,过滤式特征选择方法往往会忽略特征之间的相关性。封装式特征选择算法与后续学习算法相关,利用学习算法的性能评价所选特征子集的好坏,因此在精度方面要优于过滤式特征选择^[8-12]。基于特征选择的目的,已经有部分学者做了相关研究。例如,传统的基于空间搜索的最大相关最小冗余(minimal redundancy maximal relevance, MRMR)^[13]算法,使用互信息来度量特征之间的冗余度以及与类别之间的相关度,并且利用信息熵和信息差两个函数来选取最优特征子集。但是,由于冗余度和相关度的评价准则单一,所以使得该特征选择算法的使用范围较窄。2018年,郭凯文等^[14]提出了基于特征选择和聚类的分类算法,特征选择标准采用的是传统的基于空间搜索的最大相关最小冗余准则,将信息差作为目标函数来求解最优特征子集。虽然该算法在目标函数中增加了相关度和冗余度的权重因子,但是,在求解最优特征子集的过程中需要对权重因子不断地赋值以寻求最优子集,计算量较大;2020年,李纯果等^[15]提出的基于排序互信息的无监督特征选择,是基于排序互信息反应的两属性之间的单调关系,用每个属性与其他属性之间的平均互信息,来衡量每个属性与排序学习的相关度,平均互信息最高的视为排序最相关的属性。但是,该算法忽略了特征与特征之间的冗余度,只在低维度且样本量较少的模拟数据集上进行了有效性验证,对真实数据集的特征选择效果不明了;2020年,刘云等^[16]提出了混合蒙特卡罗搜索的特征选择算法的优化,根据蒙特卡罗树搜索方法生成了一个初始特征子集,然后利用ReliefF算法选择前 k 个特征组成候选特征集,最后,用KNN分类器的分类精度评估候选特征,选择高精度的候选特征作为最佳特征子集。然而,ReliefF算法是从同类和不同类中各选取 k 个近邻样本,求平均值得到各个特性权值,即特征与类别之间的相关性,并没有考虑特征与特征之间的冗

余度。2020年,周传华等^[17]提出的最大相关与独立分类信息最大化特征选择算法,用互信息度量特征与类别之间的相关性,用独立分类信息综合衡量新分类信息和特征冗余,尽管在特征选择过程中综合考虑了特征与类别的相关性、特征之间的冗余性,以及特征包含的新分类信息,并结合最大最小准则对特征的重要性进行了非线性评价,但其目标函数与传统的MRMR算法的目标函数类似,依然不能根据客户的实际需求进行特征选择。

针对上述特征选择算法中存在的冗余度和相关度的度量准则单一以及评价函数问题,提出了新方案。在冗余度量准则方面引入了2种不同的方法,在相关度量准则方面引入了4种不同的方法,从而组合衍生出8种特征选择算法,提出了新的目标函数。

1 新的特征选择算法

MRMR算法是最常用、最典型的基于空间搜索的特征选择算法。其中,最大相关即特征与类别间的相关度要最大,最小冗余即特征与特征之间的相关度要最小^[18-19],该算法中,冗余度和相关度均是利用互信息作为度量准则,就效能而言,比只考虑特征与类别之间的相关度,或者只考虑特征之间冗余度的特征选择算法要好。但是,在现实生活中,我们面临的数据往往纷繁复杂,面对不同的数据,MRMR算法呈现出的效果有较大差异,从而降低了该算法的适用范围。

针对MRMR算法存在的问题,提出一种新的最大相关最小冗余特征选择算法(new algorithm for feature selection with maximum relation and minimum redundancy, New-MRMR)。这里New-MRMR算法仅是新提出的一个特征选择的框架,在度量特征与特征之间冗余度时选用了2种评价准则,在度量特征与特征之间相似度时选用了4种评价准则,从而衍生出8种特征选择算法,当面对不同的用户需求时,选用不同特征选择算法,使得新提算法的适用范围更广。具体的特征选择流程见图1。

图1可以看出,特征选择算法的基本流程为:先对原始数据集进行预处理,将原始数据集分为测试集和训练集,然后,在训练集上选择不同的冗余度和相关度评价准则来训练模型,进行特征选择,得到最优特征子集,最后,利用测试集来验证模型的有效性。

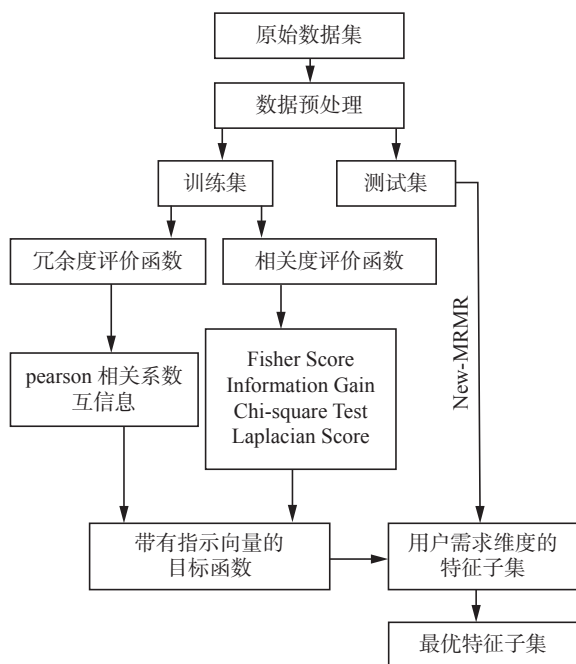


图 1 New-MRMR 特征选择流程

Fig. 1 New-MRMR feature selection flow

1.1 冗余度评价准则

特征选择是为了去除原始特征集中的冗余特征, 达到降维目的。因此, 利用冗余度评价可以作为 New-MRMR 特征选择算法的一部分, 其基本思想是: 两个特征的相关度越大, 则这两个特征冗余度也越高。但是, 由于评价特征之间冗余度以及特征与类别之间相关度的准则众多, 且目前缺乏相关研究给出具体哪种方法更适用于哪种数据类型。所以, 本文新提出的算法仅采用了 Pearson 相关系数^[14] 以及互信息^[14] 两种准则来度量特征之间的冗余度。

1.2 相关度评价准则

在特征选择过程中, 通常优先选择与类别相关度较大的特征, 而特征的重要度在一定程度上反映了与类别的相关度大小, 因此, 相关度的度量准则就转化成了特征重要度的衡量。衡量特征重要度的评价准则有很多, 例如: Fisher score^[7]、信息增益 (information gain, IG)^[8]、Laplacian Score^[20]、Chi-squar Test^[21-22] 等。Fisher score 主要是按照类内距离小, 类间距离大的原则, 选出包含鉴别信息比较多的特征, 其值越大, 说明该特征越重要, 与类别的相关度越大; 信息增益是通过计算某特征被使用前后的信息熵来为该特征进行打分, 信息增益越大, 说明该特征越重要, 与类别的相关度越大; Laplacian Score 是根据拉普拉斯特征映射等对单个特征评分, 然后选出方差和局部几何结构保持能力较强的特征, 其分值越高, 特征越重要。New-MRMR 算法也采用这 4 种评价准则作

为相关度的度量准则。

1.3 目标函数

基于特征选择和聚类的众多分类算法中, 目标函数常采用加权的信息差方式, 并且通过对权重信息不断赋值来求解最优特征子集, 不能根据不同用户实际需求的维度求解最优特征子集。因此, 本文提出了一种新的目标函数, 引入了一个指示向量 λ 以及参数 k 来表示所选的特征维度。具体目标函数如下:

$$\max_{\lambda} \left(\frac{\lambda^T D}{k} - \frac{\lambda^T C \lambda}{k(k-1)} \right), \text{ s.t. } \sum \lambda_i = k, \lambda_i \in [0, 1]$$

式中: k 为用户需求的实际数据维度; D 为冗余度矩阵; C 为特征与类别之间的相关性矩阵。 $\lambda = [\lambda_1 \lambda_2 \cdots \lambda_n]^T$, n 为原始特征集的特征数。当 λ_i 取值为 0 时, 说明对应的特征不会被选择进最终的特征子集, λ_i 取值越大时, 表明其对应的特征越容易被选进最终的特征子集。

对于该目标函数的求解, 与最优化标准二次规划问题^[23] 相似, 本文采用成对更新方法^[24] 来求解以上目标函数的最优解。

2 实验结果与分析

2.1 数据集信息及评价指标

为验证 New-MRMR 算法的有效性, 本文使用了 4 个真实的 UCI 数据集。先利用新提出的算法处理原始特征, 进而使用支持向量机对所得到的特征子集进行分类实验, 最后比较各种算法在测试集上的分类准确率 (classification precision, CP)。相关定义如下:

$$CP = \frac{CC}{Num} \times 100\%$$

式中: CC(correct classification, CC) 为正确分类的样本数量; Num 为样本数量总数。

表 1 为 4 个 UCI^[25] 数据集的具体信息:

表 1 实验数据集
Table 1 Experimental data set

数据集	特征数	样本数	测试集	训练集
isolet	617	7797	2599	5198
waveform	40	5000	1666	3334
clean	168	476	158	318
Parkinson's Disease	754	756	252	504

实验中, 与新提算法进行对比的特征选择算法分别是: Fisher Score、基于 Information Gain 的方法、基于 Laplacian Score 的方法、基于 Chi-squar Test 的方法、基于 MRMR 的方法。表 2 列出了以

上方法。

2.2 实验结果对比分析

特征选择过程是剔除原始数据集中的不相关以及冗余特征,达到数据降维目的。为验证以上

各种算法在数据降维和用支持向量机分类后的分类准确率,表3给出了以上各种算法在数据集 isolet 上的实验结果,即经支持向量机分类后,计算得到的分类准确率达到最大时所选择的特征数。

表2 新提出的8种特征选择算法与其他算法对比

Table 2 Comparison of 8 newly proposed feature selection algorithms with other algorithms

对比算法	特征选择算法	冗余度评价准则	相关度评价准则
本文新提出的8种算法	New-MRMR-F-P	Pearson相关系数	Fisher Score
	New-MRMR-F-NI	互信息	Fisher Score
	New-MRMR-L-P	Pearson相关系数	Laplacian Score
	New-MRMR-L-NI	互信息	Laplacian Score
	New-MRMR-K-P	Pearson相关系数	Chi-squar Test
	New-MRMR-K-NI	互信息	Chi-squar Test
	New-MRMR-IG-P	Pearson相关系数	Information Gian
	New-MRMR-IG-NI	互信息	Information Gian
传统算法	MRMR	互信息	互信息
	Fisher Score	Fisher Score	Fisher Score
	Laplacian Score	Laplacian Score	Laplacian Score
	Chi-squar Test	Chi-squar Test	Chi-squar Test
	Information Gian	Information Gian	Information Gian

表3 分类准确率最大时,数据集 isolet 上各种算法分别所选择的特征数

Table 3 Number of features selected by various algorithms when the Classification precision is maximum on the isolet data-set

对比算法	特征选择算法	所选特征数	分类准确率
本文新提出的8种算法	New-MRMR-F-P	340	0.958 3
	New-MRMR-F-NI	341	0.940 5
	New-MRMR-IG-P	343	0.963 5
	New-MRMR-IG-NI	387	0.958 3
	New-MRMR-K-P	289	0.955 8
	New-MRMR-K-NI	332	0.955 7
	New-MRMR-L-P	342	0.945 7
	New-MRMR-L-NI	288	0.945 3
传统算法	MRMR	391	0.908 8
	Fisher Score	389	0.885 7
	Laplacian Score	487	0.918 4
	Chi-square Test	542	0.933 5
	Information Gain	489	0.916 1

由表3可以看出,由以上各种算法对数据集 isolet 进行特征选择后,利用支持向量机对所选特征子集进行分类,本文新提出的8种特征选择算法的分类准确率,均高于传统的5种特征选择算

法,尤其是新提出的算法 New-MRMR-IG-P,其分类准确率达到0.963 5,远高于传统的5种特征选择算法。在保证准确率的情况下,其所选的特征数也均小于传统的5种特征选择算法。可见,本文

新提出的特征选择算法在数据降维方面效果更佳。

图2是在数据集 isolet 上,本文新提出的特征选择算法 New-MRMR-F-NI、New-MRMR-F-P,传统特征选择算法 MRMR、Fisher Score 在不同维度下的分类准确率变化趋势。

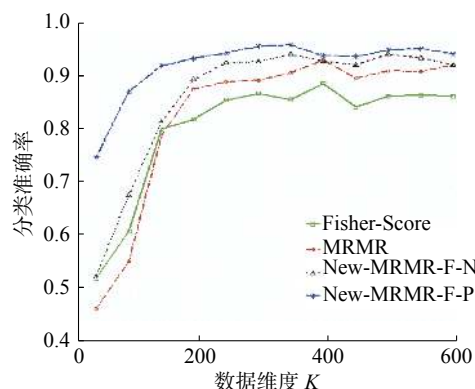


图2 New-MRMR-F-NI、New-MRMR-F-P、Fisher-Score、MRMR 在数据集 isolet 上分类准确率的变化趋势
Fig. 2 Correct classification trend of New-MRMR- F-NI, New-MRMR-F-P, Fisher-Score, MRMR on the dataset isolet

从图2可以看出,对于在不同维度下的分类准确率,新提出的特征选择算法 New-MRMR-F-NI、New-MRMR-F-P 明显高于传统算法 Fisher Score、MRMR。所以,对于减少原始特征集中的冗余和不相关特征,New-MRMR-F-NI、New-MRMR-F-P 有更好的优势。

不同维度下,本文新提算法 New-MRMR-K-NI、New-MRMR-K-P,传统算法 MRMR、Chi-Square-Test 在数据集 isolet 上的分类准确率变化趋势见图3。

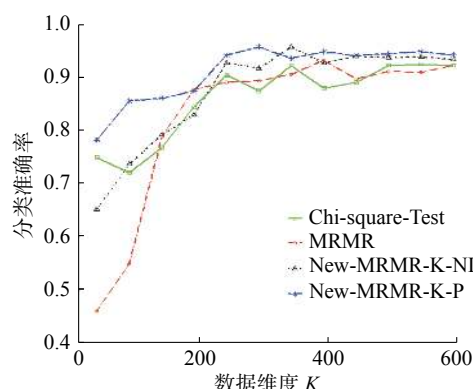


图3 New-MRMR-K-NI、New-MRMR-K-P、Chi-Square-Test、MRMR 在数据集 isolet 上分类准确率的变化趋势
Fig. 3 Correct classification trend of New-MRMR-K-NI, New-MRMR-K-P, Chi-Square-Test, MRMR on the dataset isolet

图3显示,不同维度下,New-MRMR-K-P 的分类准确率曲线明显高于传统特征选择算法,并

且,在所选特征子集数为 289 时,其分类准确率达到最高,既很好地去除了原始特征集中的冗余和不相关特征,又保证了分类准确率。此外,算法 New-MRMR-K-P 除了在维度为 195 时的分类准确率与传统算法 MRMR 相近之外,在其他维度上的分类准确率均高于 Chi-Square-Test、MRMR。可见,本文新提出的特征选择算法效果更佳。

不同维度下,新提出的特征选择算法 New-MRMR-L-NI、New-MRMR-L-P,传统特征选择算法 MRMR、Laplacian-Score 的分类准确率变化趋势见图4。

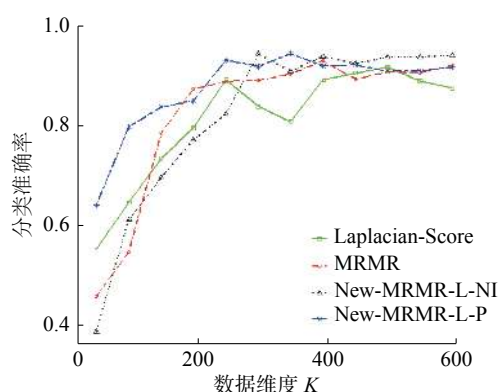


图4 New-MRMR-L-NI、New-MRMR-L-P、Laplacian-Score、MRMR 在数据集 isolet 上,分类正确的变化趋势

Fig. 4 Correct classification trend of New-MRMR-L-NI, New-MRMR-L-P, Laplacian-Score, MRMR on the dataset isolet

图4显示,在特征维度为 342 的时候,算法 New-MRMR-L-P 的分类准确率就已经达到了最高,并且大于传统算法 Laplacian-Score、MRMR 的最大分类准确率。此外,在分类准确率达到最高时,算法 New-MRMR-L-NI 所选的特征子集数仅为 288,远小于传统算法 Laplacian-Score、MRMR 所选的特征子集数。因此,新提出的算法 New-MRMR-L-NI、New-MRMR-L-P 对于特征选择效果更好。

不同维度下,新提出的特征选择算法 New-MRMR-IG-NI、New-MRMR-IG-P,传统特征选择算法 MRMR、Laplacian-Score 的分类准确率变化趋势见图5。

由图5可以看出,在不同维度下,算法 New-MRMR-IG-NI、New-MRMR-IG-P 分类准确率的曲线,均高于传统的两种特征选择算法 Information-Gain、MRMR 所代表的曲线。分类准确率越高,表明所选特征子集越好。可见,新出的算法 New-MRMR-IG-NI 以及 New-MRMR-IG-P 在特征选择方面更加有效。

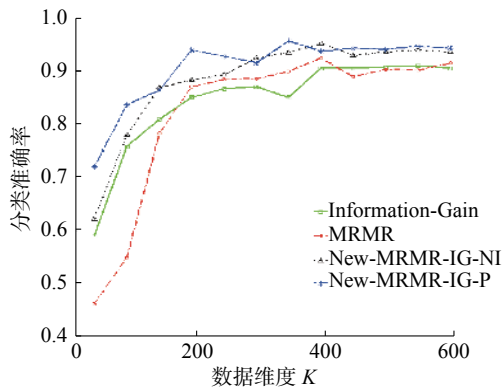


图5 New-MRMR-IG-NI、New-MRMR-IG-P、Information-Gain、MRMR 在数据集 isolet 上, 分类准确率的变化趋势

Fig. 5 Correct classification trend of New-MRMR-IG-NI, New-MRMR-IG-P, Information-Gain, MRMR on the dataset isolet

表4 分类准确率最大时数据集 waveform 上各种算法分别所选择的特征数

Table 4 Number of features selected by various algorithms when the Classification precision is maximum on the waveform dataset

对比算法	特征选择算法	所选特征数	分类准确率
本文提出的8种算法	New-MRMR-F-P	16	0.953 4
	New-MRMR-F-NI	20	0.890 4
	New-MRMR-IG-P	24	0.891 6
	New-MRMR-IG-NI	17	0.900 0
	New-MRMR-K-P	12	0.941 2
	New-MRMR-K-NI	20	0.894 4
	New-MRMR-L-P	17	0.889 2
	New-MRMR-L-NI	21	0.934 4
传统算法	MRMR	20	0.862 4
	Fisher Score	40	0.838 4
	Laplacian Score	25	0.725 2
	Chi-square Test	32	0.809 5
	Information Gain	33	0.798 3

不同维度下, 本文新提出的特征选择算法 New-MRMR-F-NI、New-MRMR-F-P, 传统特征选择算法 MRMR、Fisher Score 在数据集 waveform 上的分类准确率变化趋势见图6。

由图6看出, 在数据集 waveform 上, New-MRMR-F-P 的表现最好, 其所代表的曲线远高于传统的特征选择算法 MRMR、Fisher-Score 所代表的曲线。此外, 虽然在维度为24时, 算法 New-MRMR-F-NI 的分类准确率低于传统算法 MRMR、Fisher-Score。但是, 在其余维度上, New-MRMR-F-NI 的分类准确率均高于 MRMR、Fisher-

Score。综合分析, 本文新提算法 New-MRMR-F-NI、New-MRMR-F-P 的特征选择效果更好。

不同维度下, 算法 New-MRMR-K-NI、New-MRMR-K-P 以及传统特征选择算法 MRMR 以及 Chi-Square-Test 在数据集 waveform 上的分类准确率变化趋势见图7。

图7显示, 维度为20时, New-MRMR-K-NI 的分类准确率就达到了最大, 大于 MRMR、Chi-Square-Test 的最大分类准确率。并且其所选特征子集数小于 MRMR、Chi-Square-Test 的最优特征子集数。此外, 算法 New-MRMR-K-P 的分类准确

率曲线高于 MRMR、Chi-Square-Test 的分类准确率曲线。所以,在 waveform 数据集上,本文新提出的算法 New-MRMR-K-NI、New-MRMR-K-P 的特征选择效果更好。

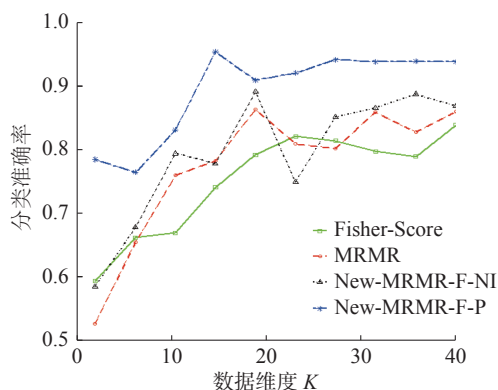


图6 New-MRMR-F-NI、New-MRMR-F-P、Fisher-Score、MRMR 在数据集 waveform 上,分类准确率的变化趋势

Fig. 6 Correct classification trend of New-MRMR-F-NI, New-MRMR-F-P, Fisher-Score, MRMR on the dataset waveform

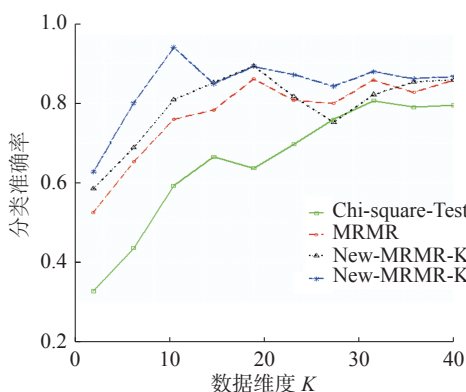


图7 New-MRMR-K-NI、New-MRMR-K-P、Chi-Square-Test、MRMR 在数据集 waveform 上,分类准确率的变化趋势

Fig. 7 Correct classification trend of New-MRMR-K-NI, New-MRMR-K-P, Chi-Square-Test, MRMR on the dataset waveform

不同维度下,算法 New-MRMR-L-NI、New-MRMR-L-P,传统特征选择算法 MRMR、Laplacian-Score 在数据集 waveform 上的分类准确率变化趋势见图 8。

图 8 显示, New-MRMR-L-NI 的分类准确率高出传统算法 MRMR、Laplacian-Score。在分类准确率达到最大时, New-MRMR-L-NI 所选特征子集数仅为 20, 小于 MRMR、Laplacian-Score 的最优特征子集数。另外,新提算法在多数维度上均大于传统算法 MRMR、Laplacian-Score 的分类准确率。由于分类准确率越高,特征选择效果越好,所以,在数据集 waveform 上, New-MRMR-L-NI、

New-MRMR-L-P 的特征选择效果更好。

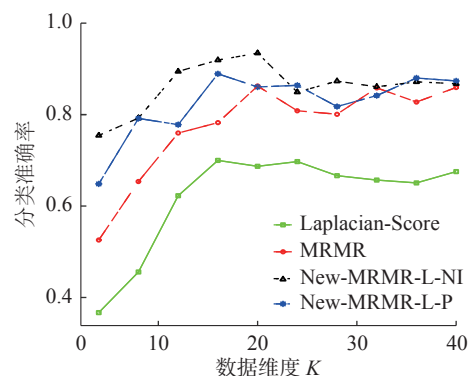


图8 New-MRMR-L-NI、New-MRMR-L-P、Laplacian-Score、MRMR 在数据集 waveform 上,分类准确率的变化趋势

Fig. 8 Correct classification trend of New-MRMR-L-NI, New-MRMR-L-P, Laplacian-Score, MRMR on the dataset waveform

不同维度下, New-MRMR-IG-NI、New-MRMR-IG-P、传统算法 MRMR、Information-Gain 在数据集 waveform 上分类准确率变化趋势见图 9。

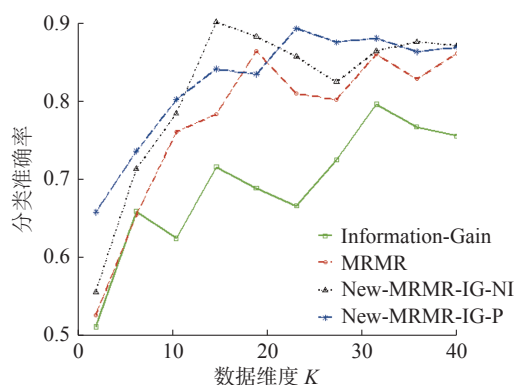


图9 New-MRMR-IG-NI、New-MRMR-IG-P、Information-Gain、MRMR 在数据集 waveform 上,分类准确率的变化趋势

Fig. 9 Correct classification trend of New-MRMR-IG-NI, New-MRMR-IG-P, Information-Gain, MRMR on the dataset waveform

图 9 显示,在数据集 waveform 上,算法 New-MRMR-IG-NI 的分类准确率的曲线高于传统的算法 MRMR、Information-Gain 的分类准确率。且算法 New-MRMR-IG-P 的分类准确率在维度为 24 时达到最大。维度为 11 时, New-MRMR-IG-P 的分类准确率略低于 MRMR、Information-Gain,但是,在其余维度上均大于 MRMR、Information-Gain。综上分析,在数据集 waveform 上,本文新提出的特征选择算法效果明显。

表 5 给出了以上各种算法在数据集 clean 上的实验结果,即经支持向量机分类后,得到的分类准确率达到最大时所选择的特征数。

表5 分类准确率最大时数据集 clean 上各种算法分别所选择的特征数

Table 5 Number of features selected by various algorithms when the Classification precision is maximum on the clean data-set

对比算法	特征选择算法	所选特征数	分类准确率
本文提出的8种算法	New-MRMR-F-P	50	0.865 6
	New-MRMR-F-NI	70	0.865 5
	New-MRMR-IG-P	130	0.865 4
	New-MRMR-IG-NI	60	0.848 7
	New-MRMR-K-P	130	0.865 8
	New-MRMR-K-NI	20	0.875 9
	New-MRMR-L-P	90	0.827 3
	New-MRMR-L-NI	40	0.883 7
传统算法	MRMR	110	0.817 5
	Fisher Score	90	0.798 2
	Laplacian Score	140	0.782 3
	Chi-square Test	130	0.808 4
	Information Gain	130	0.825 9

由表5可以看出,在分类准确率方面,本文新提出的算法的最大分类准确率均高于5种传统的特征选择算法。在分类准确率达到最优时所选的特征子集数方面,尤其是算法 New-MRMR-K-NI,其所选的特征子集数仅20,远小于原始的特征子集数。所以,对于数据集 clean 而言,本文新提出的特征选择算法更加有效。

不同维度下,算法 New-MRMR-F-NI、New-MRMR-F-P、传统特征选择算法 MRMR、Fisher Score 在数据集 clean 上的分类准确率变化趋势见图10。

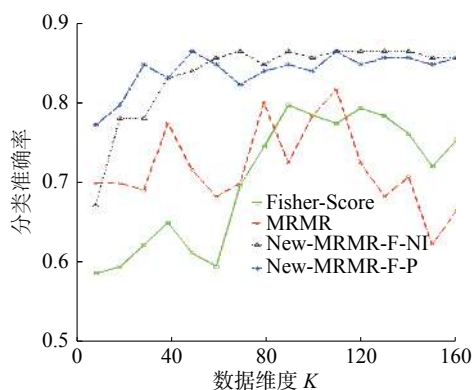


图10 New-MRMR-F-NI、New-MRMR-F-P、Fisher-Score、MRMR 在数据集 clean 上分类准确率的变化趋势

Fig. 10 Correct classification trend of New-MRMR-F-NI, New-MRMR-F-P, Fisher-Score, MRMR on the dataset clean

由图10可以看出,本文新提算法 New-MRMR-F-NI、New-MRMR-F-P 的分类准确率曲线均 MRMR、Fisher-Score 的分类准确率的曲线之上。由此可见,在数据集 clean 上,算法 New-MRMR-F-NI、New-MRMR-F-P 的特征选择结果更优。

不同维度下,算法 New-MRMR-K-NI、New-MRMR-K-P、传统特征选择算法 MRMR、Chi-Square-Test 在数据集 clean 上的分类准确率变化趋势见图11。

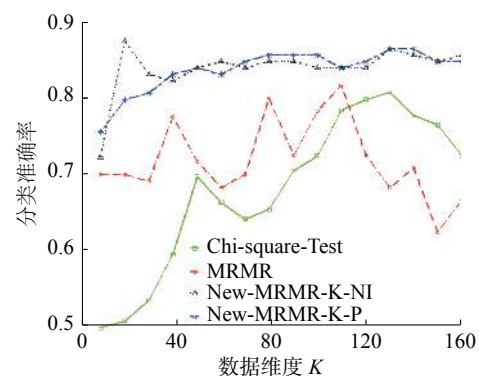


图11 New-MRMR-K-NI、New-MRMR-K-P、Chi-Square-Test、MRMR 在数据集 clean 上, 分类准确率的变化趋势

Fig. 11 Correct classification trend of New-MRMR-K-NI, New-MRMR-K-P, Chi-Square-Test, MRMR on the dataset clean

图11中, New-MRMR-K-NI、New-MRMR-K-P 的分类准确率的曲线均在传统的特征选择算法

MRMR、Chi-quare-Test 之上,尤其是 New-MRMR-K-NI,当分类准确率达到最大时,所选的特征子集数为 20,远小于两种传统算法所选择的最优特征子集数。可见,在数据集 clean 上,算法 New-MRMR-K-NI、New-MRMR-K-P 的特征选择效果更优。

不同维度下,算法 New-MRMR-L-NI、New-MRMR-L-P、传统特征选择算法 MRMR、Fisher Score 在数据集 clean 上的分类准确率变化趋势见图 12。

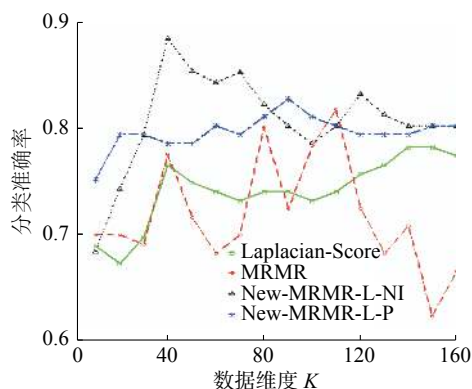


图 12 New-MRMR-L-NI、New-MRMR-L-P、Laplacian-Score、MRMR 在数据集 clean 上分类准确率的变化趋势

Fig. 12 Correct classification trend of New-MRMR-L-NI, New-MRMR-L-P, Laplacian-Score, MRMR on the dataset clean

图 12 可以看出,维度为 40 时,算法 New-MRMR-L-NI 就达到了最大分类准确率,且高于传统算法 MRMR、Laplacian-Score 的分类准确率。此外,虽然在维度为 110 时,New-MRMR-L-P 的分类准确率略低于 MRMR,但在其余维度上的分类准确率均高于 MRMR、Laplacian-Score 的分类准确率。

可见,在数据集 clean 上,新提算法 New-MRMR-L-NI、New-MRMR-L-P 的特征选择效果更好。

不同维度下,算法 New-MRMR-IG-NI、New-MRMR-IG-P、传统特征选择算法 MRMR、Fisher Score 在数据集 clean 上的分类准确率变化趋势见图 13。

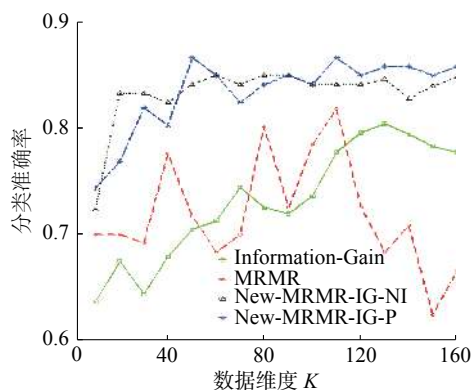


图 13 New-MRMR-IG-NI、New-MRMR-IG-P、Information-Gain、MRMR 在数据集 clean 上分类准确率的变化趋势

Fig. 13 Correct classification trend of New-MRMR-IG-NI, New-MRMR-IG-P, Information-Gain, MRMR on the dataset clean

图 13 显示,本文新提算法 New-MRMR-IG-NI、New-MRMR-IG-P 的分类准确率曲线均在传统算法的分类准确率曲线之上。所以,对于数据集 clean,本文新提出的两种特征选择算法 New-MRMR-IG-NI、New-MRMR-IG-P 所选择的特征子集更加有效。

表 6 给出了以上各种算法在数据集 Parkinson's Disease 上的实验结果,即经支持向量机分类后,得到的分类准确率达到最大时所选择的特征数。

表 6 分类准确率最大时,数据集 Parkinson's Disease 上各种算法分别所选择的特征数

Table 6 Number of features selected by various algorithms when the Classification precision is maximum on the Parkinson's Disease dataset

对比算法	特征选择算法	所选特征数	分类准确率
本文新提出的8种算法	New-MRMR-F-P	150	0.914 2
	New-MRMR-F-NI	210	0.877 9
	New-MRMR-IG-P	180	0.883 4
	New-MRMR-IG-NI	210	0.914 2
	New-MRMR-K-P	300	0.862 4
	New-MRMR-K-NI	120	0.887 4
	New-MRMR-L-P	150	0.881 6
	New-MRMR-L-NI	240	0.901 6
传统算法	MRMR	540	0.867 0
	Fisher Score	540	0.825 3

续表 6

对比算法	特征选择算法	所选特征数	分类准确率
	Laplacian Score	240	0.810 0
	Chi-square Test	300	0.724 9
	Information Gain	210	0.753 8

表 6 显示, 算法 New-MRMR-F-P 的分类准确率高达 0.9124, 且此时所选择的特征子集数仅为 150, 远小于传统的 5 种算法的最优特征子集数。另外, 除了 New-MRMR-K-P 的分类准确率略低于传统算法 MRMR 的分类准确率之外, 新提出的其余算法均大于传统特征选择算法。由此可见, 本文新提出的特征选择算法在数据集 Parkinson's Disease 上的特征选择效果更好。

不同维度下, 算法 New-MRMR-F-NI、New-MRMR-F-P, 传统特征选择算法 MRMR、Fisher-Score 在数据集 Parkinson's Disease 上的分类准确率变化趋势见图 14。

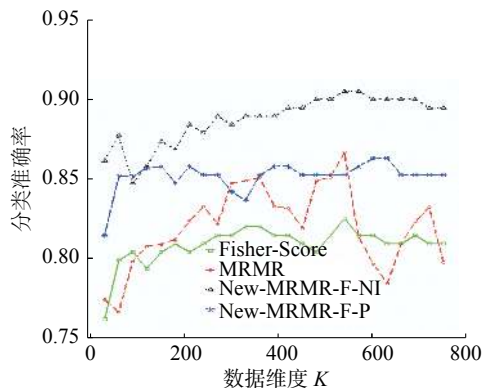


图 14 New-MRMR-F-NI、New-MRMR-F-P、Fisher-Score、MRMR 在数据集 Parkinson's Disease 上分类准确率的变化趋势

Fig. 14 Correct classification trend of New-MRMR-F-NI, New-MRMR-F-P, Fisher-Score, MRMR on the Parkinson's Disease dataset

图 14 显示, 算法 New-MRMR-F-NI 的分类准确率曲线在传统算法 MRMR、Fisher-Score 的分类准确率曲线之上。在维度为 540 时, New-MRMR-F-P 的分类准确率略低于 MRMR 的分类准确率。但是, 在其余维度上, New-MRMR-F-P 的分类准确率均高于传统算法 MRMR、Fisher-Score 的分类准确率。更重要的是, 在达到最大分类准确率时, New-MRMR-F-NI 所选的特征子集数仅为 210, 远低于 MRMR、Fisher-Score 的最优特征子集数。所以, 在数据集 Parkinson's Disease 上, 本文新提出的算法特征选择效果更好。

不同维度下, 本文新提算法 New-MRMR-F-NI、New-MRMR-F-P、传统算法 MRMR、Fisher-

Score 在数据集 Parkinson's Disease 上的分类准确率变化趋势见图 15。

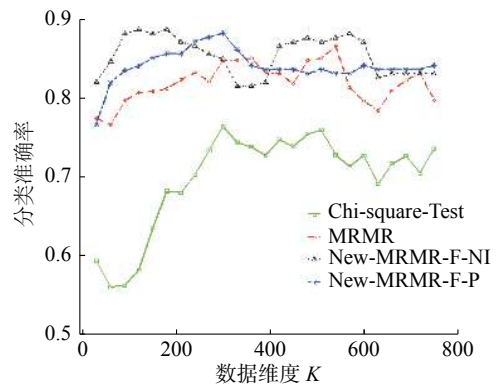


图 15 New-MRMR-K-NI、New-MRMR-K-P、Chi-Square-Test、MRMR 在数据集 Parkinson's Disease 上分类准确率的变化趋势

Fig. 15 Correct classification trend of New-MRMR-K-NI, New-MRMR-K-P, Chi-Square-Test, MRMR on the Parkinson's Disease dataset

由图 15 可见, 在绝大多数维度上, New-MRMR-F-NI、New-MRMR-F-P 的分类准确率均高于 MRMR、Chi-Square-Test 的分类准确率。在维度为 120 时, New-MRMR-F-NI 就已然达到了最大分类准确率, 大于 MRMR、Chi-Square-Test 的最大分类准确率。由此可见, 在数据集 Parkinson's Disease 上, 本文新提算法特征选择效果更好。

不同维度下, 本文新提出的特征选择算法 New-MRMR-F-NI、New-MRMR-F-P 以及传统特征选择算法 MRMR 以及 Fisher Score 在数据集 Parkinson's Disease 上的分类准确率变化趋势见图 16。

由图 16 可以看出, 算法 New-MRMR-L-P 的分类准确率的曲线高于传统算法 MRMR、Laplacian-Score 的分类准确率曲线, 并且, 在维度为 240 时, New-MRMR-L-NI 就已经达到了最大分类准确率, 远小于 MRMR 达到最大分类准确率时所选择的特征子集数 (540)。由此可见, 在数据集 Parkinson's Disease 上, 本文新提算法特征选择效果更好。

不同维度下, 本文新提算法 New-MRMR-F-NI、New-MRMR-F-P 以及传统算法 MRMR、Fisher Score 在数据集 Parkinson's Disease 上的分类准确率变化趋势见图 17。

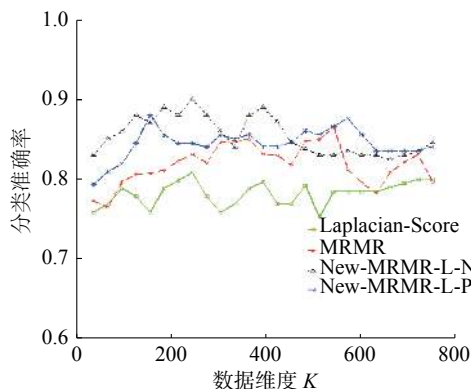


图 16 New-MRMR-L-NI、New-MRMR-L-P、Laplacian-Score、MRMR 在数据集 Parkinson's Disease 上分类准确率的变化趋势

Fig. 16 Correct classification trend of New-MRMR-L-NI, New-MRMR-L-P, Laplacian-Score, MRMR on the Parkinson's Disease dataset

由图 17 可以看出,在维度为 120 和 540 时, New-MRMR-IG-P 的分类准确率与算法 MRMR 的分类准确率较为接近,但在其余维度上,其分类准确率均大于 MRMR 的分类准确率。而且,在分类准确率达到最大时, New-MRMR-IG-P 所选择的特征子集数仅为 180,远小于 MRMR 的最优特征子集数。此外, New-MRMR-IG-NI 的分类准确率的曲线高于算法 MRMR、Information-Gain 的分类准确率曲线。由上述分析可知,针对数据集 Parkinson's Disease 而言,本文提出算法在整体

上比传统算法选择结果更好。

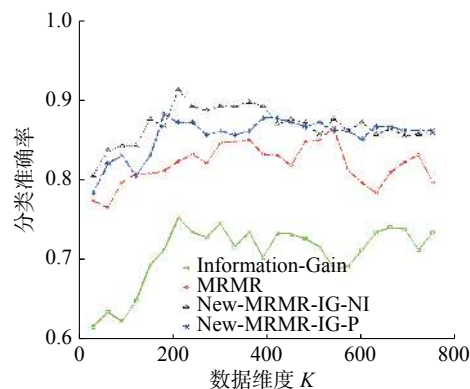


图 17 New-MRMR-IG-NI、New-MRMR-IG-P、Information-Gain、MRMR 在数据集 Parkinson's Disease 上分类准确率的变化趋势

Fig. 17 Correct classification trend of New-MRMR-IG-NI, New-MRMR-IG-P, Information-Gain, MRMR on the Parkinson's Disease dataset

2.3 实验结果的 T 检验

为更加有效地证明本文新提的 8 种特征选择算法的有效性,以下采用成对单边 T 检验来证明其有效性。原假设为:本文新提算法与传统算法的特征选择效果相同;备择假设为:本文新提算法的特征选择效果优于传统特征选择算法。表 7 为假设检验结果,其中包含了检验的统计量,置信区间以及 P 值。

表 7 新提算法与传统算法的成对单边 T 检验的检验结果

Table 7 Test results of paired unilateral T-test between the new algorithm and the traditional algorithm

成对T检验	P值	T统计量	95%置信区间
New-MRMR-F-NI vs MRMR	0.0268	3.0890	(0.7987, Inf)
New-MRMR-F-P vs MRMR	0.0088	4.7566	(0.8302, Inf)
New-MRMR-K-NI vs MRMR	0.0271	3.0765	(0.7966, Inf)
New-MRMR-K-P vs MRMR	0.0152	3.8754	(0.8126, Inf)
New-MRMR-L-NI vs MRMR	0.0105	4.4561	(0.8230, Inf)
New-MRMR-L-P vs MRMR	0.0321	2.8649	(0.7933, Inf)
New-MRMR-IG-NI vs MRMR	0.0206	3.4353	(0.8078, Inf)
New-MRMR-IG-P vs-MRMR	0.0189	3.5595	(0.8089, Inf)
New-MRMR-F-NI vs Fisher-Score	0.0134	4.0658	(0.7987, Inf)
New-MRMR-F-P vs Fisher-Score	0.0005	5.7347	(0.8302, Inf)
New-MRMR-K-NI vs Chi-Square-Test	0.0030	6.9165	(0.7966, Inf)
New-MRMR-K-P vs Chi-Square-Test	0.0026	7.3174	(0.8126, Inf)
New-MRMR-L-NI vs Laplacian-Score	0.0328	2.8405	(0.7730, Inf)
New-MRMR-L-P vs Laplacian-Score	0.0130	4.1147	(0.7933, Inf)
New-MRMR-IG-NI vs Information-Gain	0.0064	5.3303	(0.8068, Inf)
New-MRMR-IG-P vs Information-Gain	0.0056	5.5739	(0.8089, Inf)

由表7可以看出,成对单边T检验的 P 值均小于0.05,所以拒绝原假设,故认为本文新提出的8种特征选择算法的特征选择结果优于传统特征选择算法的特征选择结果。

综上所述,从分类准确率以及假设检验的结果可以看出,本文新提出的8种特征选择算法所选择的特征子集更优,特征选择效果更好。

3 结束语

虽然传统的基于特征选择的分类算法的理念已较为新颖,但是还是存在一定的提升空间。一方面,传统的基于特征选择的分类算法在特征选择过程中采用的度量特征之间冗余度以及和类别的相关度的评价准则单一;另一方面,它只考虑了特征与类别之间的相关度而忽略了冗余度;最后,其目标函数也存在缺陷,不能根据用户实际的维度需求来选择特征子集。本文针对这些问题引入了4种不同的相关度评价准则以及两种不同的冗余度评价准则,目标函数中引入了指示向量 λ 来刻画用户实际的数据维度需求,从而组合成8种新的特征选择算法,利用支持向量机对这8种算法选择得到的特征子集分类。在4个真实的UCI数据集上进行了实验,利用分类准确率和T检验验证了新提出的算法的有效性。

最后需要指出,评价特征冗余度和相关度的方法有多种,本文仅用了2种评价冗余度的方法和4种评价相关度的方法,但是其他评价冗余度和相关度的方法也可以适用于New-MRMR框架,此外,新提特征选择算法在不同数据集上表现性能不同。因此,后续研究中,会更深入地研究和挖掘数据本质,尝试利用足够多的数据集以及评价相关度和冗余度的方法来深入探索具体哪种算法更适合哪种领域。

参考文献:

- [1] 王娟,慈林林,姚康泽. 特征选择方法综述[J]. 计算机工程与科学, 2005, 27(12): 68–71.
WANG Juan, CI Linlin, YAO Kangze. A survey of feature selection[J]. Computer engineering and science, 2005, 27(12): 68–71.
- [2] 周红标,乔俊飞. 基于高维 k -近邻互信息的特征选择方法[J]. 智能系统学报, 2017, 12(5): 595–600.
ZHOU Hongbiao, QIAO Junfei. Feature selection method based on high dimensional k -nearest neighbors mutual information[J]. CAAI transactions on intelligent systems, 2017, 12(5): 595–600.
- [3] 徐洪峰,孙振强. 多标签学习中基于互信息的快速特征选择方法[J]. 计算机应用, 2019, 39(10): 2815–2821.
XU Hongfeng, SUN Zhenqiang. Fast feature selection method based on mutual information in multi-label learning[J]. Journal of computer applications, 2019, 39(10): 2815–2821.
- [4] 计智伟,胡珉. 一种双重过滤式特征选择算法[J]. 计算机工程与应用, 2011, 47(19): 190–193, 206.
JI Zhiwei, HU Min. A double-filtered feature selection algorithm[J]. Computer engineering and applications, 2011, 47(19): 190–193, 206.
- [5] 吴红霞,吴悦,刘宗田,等. 基于Relief和SVM-RFE的组合式SNP特征选择[J]. 计算机应用研究, 2012, 29(6): 2074–2077.
WU Hongxia, WU Yue, LIU Zongtian, et al. Combined SNP feature selection based on Relief and SVM-RFE[J]. Application research of computers, 2012, 29(6): 2074–2077.
- [6] 肖忆南,谢榕,杜娟. 基于 t 检验和弹性网的数据分类特征选择方法[J]. 小型微型计算机系统, 2015, 36(10): 2213–2217.
XIAO Yinan, XIE Rong, DU Juan. Feature selection method for data classification based on t -test and elastic net[J]. Journal of Chinese computer systems, 2015, 36(10): 2213–2217.
- [7] SRIVIDHYA S, MALLIKA R. Feature selection for high dimensional imbalanced datasets using game theory and fisher score[J]. Journal of advanced research in dynamical and control systems, 2017, 9(8): 195–202.
- [8] SAINZ L V, SAINZ J S, LÁZARO M. Oscillatory brain activity in morphological parsing of complex words[J]. Information gain from stems and suffixes, 2018, 9(5): 271.
- [9] MALDONADO S, WEBER R. A wrapper method for feature selection using support vector machines[J]. Information sciences, 2009, 179(13): 2208–2217.
- [10] 林棋,张宏,李千目. 一种基于MA-LSSVM的封装式特征选择算法[J]. 南京理工大学学报, 2016, 40(1): 10–16.
LIN Qi, ZHANG Hong, LI Qianmu. Wrapper feature selection algorithm based on MA-LSSVM[J]. Journal of Nanjing University of Science and Technology, 2016, 40(1): 10–16.
- [11] 胡峰,杨梦. 基于特征聚类的封装特征选择算法[J]. 计算机工程与设计, 2018, 39(1): 230–237.
HU Feng, YANG Meng. Algorithm for wrapper feature selection based on feature clustering[J]. Computer engineering and design, 2018, 39(1): 230–237.
- [12] 王晓初,王士同,包芳,等. 最小化类内距离和分类算法[J]. 电子与信息学报, 2016, 38(3): 532–540.
WANG Xiaochu, WANG Shitong, BAO Fang, et al. Minimize intra-class distance and classification algorithm[J]. Journal of electronics & information technology, 2016,

- 38(3): 532–540.
- [13] PENG Hanchuan, LONG Fuhui, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2005, 27(8): 1226–1238.
- [14] 郭凯文, 潘宏亮, 侯阿临. 基于特征选择和聚类的分类算法[J]. *吉林大学学报(理学版)*, 2018, 56(2): 395–398.
GUO Kaiwen, PAN Hongliang, HOU Alin. Classification algorithm based on feature selection and clustering[J]. *Journal of Jilin University (science edition)*, 2018, 56(2): 395–398.
- [15] 李纯果, 张春琴, 李海峰. 基于排序互信息的无监督特征选择[J]. *河北大学学报(自然科学版)*, 2020, 40(2): 200–204.
LI Chunguo, ZHANG Chunqin, LI Haifeng. Unsupervised feature selection based on ranking mutual information[J]. *Journal of Hebei University (natural science edition)*, 2020, 40(2): 200–204.
- [16] 刘云, 肖雪, 黄荣乘. 混合蒙特卡罗搜索的特征选择算法的优化[J]. *信息技术*, 2020, 44(5): 28–31, 36.
LIU Yun, XIAO Xue, HUANG Rongcheng. Optimization of feature selection based on hybrid Monte Carlo Tree[J]. *Information technology*, 2020, 44(5): 28–31, 36.
- [17] 周传华, 李鸣, 吴幸运. 最大相关与独立分类信息最大化特征选择算法[J]. *计算机技术与发展*, 2020, 30(8): 46–52.
ZHOU Chuanhua, LI Ming, WU Xingyun. Feature selection algorithm for maximizing maximum correlation and independent classification information[J]. *Computer technology and development*, 2020, 30(8): 46–52.
- [18] 张俐, 王枏. 基于最大相关最小冗余联合互信息的多标签特征选择算法[J]. *通信学报*, 2018, 39(5): 111–122.
ZHANG Li, WANG Cong. Multi-label feature selection algorithm based on joint mutual information of max-relevance and min-redundancy[J]. *Journal on communications*, 2018, 39(5): 111–122.
- [19] 李扬, 顾雪平. 基于改进最大相关最小冗余判据的暂态稳定评估特征选择[J]. *中国电机工程学报*, 2013, 33(34): 179–186.
LI Yang, GU Xueping. Feature selection for transient stability assessment based on improved maximal relevance and minimal redundancy criterion[J]. *Proceedings of the CSEE*, 2013, 33(34): 179–186.
- [20] 胡敏杰, 林耀进, 王晨曦, 等. 基于拉普拉斯评分的多标记特征选择算法[J]. *计算机应用*, 2018, 38(11): 3167–3174.
- HU Minjie, LIN Yaojin, WANG Chenxi, et al. Multi-label feature selection algorithm based on Laplace score[J]. *Journal of computer applications*, 2018, 38(11): 3167–3174.
- [21] 陈湛, 梁雪春. 基于基尼指标和卡方检验的特征选择方法[J]. *计算机工程与设计*, 2019, 40(8): 2342–2345, 2360.
CHEN Chen, LIANG Xuechun. Feature selection method based on Gini index and chi-square test[J]. *Computer engineering and design*, 2019, 40(8): 2342–2345, 2360.
- [22] 徐明, 高翔, 许志刚, 等. 基于改进卡方统计的微博特征提取方法[J]. *计算机工程与应用*, 2014, 50(19): 113–117, 142.
XU Ming, GAO Xiang, XU Zhigang, et al. Microblog feature extraction method based on improved chi-square statistics[J]. *Computer engineering and applications*, 2014, 50(19): 113–117, 142.
- [23] BOMZE I M, DE KLERK E. Solving standard quadratic optimization problems via linear, semidefinite and copositive programming[J]. *Journal of global optimization*, 2002, 24(2): 163–185.
- [24] LIU Hairong, YANG Xingwei, LATECKI L J, et al. Dense neighborhoods on affinity graph[J]. *International journal of computer vision*, 2012, 98(1): 65–82.
- [25] DORA S, SUBRAMANIAN K, SURESH S, et al. Development of a self-regulating evolving spiking neural network for classification problem[J]. *Neurocomputing*, 2016, 171: 1216–1229.

作者简介:



李顺勇, 教授, 博士, 主要研究方向为统计机器学习。发表学术论文30余篇。



王改变, 硕士研究生, 主要研究方向为统计机器学习。