基因数据的交互依赖特征选择算法

张 俐*

(江苏理工学院计算机工程学院 江苏 常州 213001)

【摘要】特征选择是生物信息领域中数据预处理阶段必不可少的步骤。传统特征选择算法忽视了特征之间的依赖相关性和冗余性，因此提出一种联合互信息的特征选择算法 (JFRR)。该算法利用互信息计算特征之间的冗余值，并利用联合互信息分别计算已选特征集合、候选特征及类标签之间的相关性。将 JFRR 与其他 6 个特征选择算法在 2 个分类器上，使用 9 个不同基因数据集，进行分类准确率指标 (Precision_micro 和 F1_micro) 验证。实验结果表明，该算法能有效提高分类精度。

关 键 词 分类； 特征选择； 联合互信息； 互信息； 相关性

中图分类号 TP181 **文献标志码** A doi:10.12178/1001-0548.2021136

An Algorithm for Cross-Dependent Feature Selection of Genetic Data

ZHANG Li*

(College of Computer Engineering, Jiangsu University of Technology Changzhou Jiangsu 213001)

Abstract Feature selection is an essential step in the data preprocessing phase in the field of bioinformatics. Traditional feature selection algorithms ignore the problems of dependency relevance and redundancy between features. This paper proposes a joint feature relevance and redundancy (JFRR) algorithm for feature selection. The algorithm uses mutual information to calculate the redundancy values between features and applies joint mutual information to compute the relevance among the set of selected features, candidate features and class labels. Finally, JFRR is validated with the other six feature selection algorithms on two classifiers using nine different gene datasets with classification accuracy metrics (Precision_micro and F1_micro). The experimental results show that the JFRR method can effectively improve classification accuracy.

Key words classification; feature selection; joint mutual information; mutual information; relevance

过去几十年，在生物信息领域产出大量基因数据[1-2]。这些基因数据普遍具有样本小、维度高和高噪声等特点[3]。如何处理这些不相关和冗余特征给数据降维带来重大挑战。常见的数据降维包括特征提取[4] 和特征选择[5] 两类。特征选择由于可以删除无关和冗余特征，同时保留相关原始特征，因此引起许多关注。

在特征选择中主要有数据层面 (过滤式方法) 和算法层面 (包装器方法和嵌入式方法)[6-8] 两方面的研究。过滤式特征选择算法凭借其计算成本低、与具体分类器分离及应用领域广等优点，逐渐成为特征选择技术中的研究热点。常见的基于信息论的过滤式特征选择算法包括采用平均冗余策略的特征选择算法 (MID[9]、MIQ[9]、JMI[10] 和 CFR[11] 等) 和采用“最大最小”极端标准的特征选择算法 (CMIM[12]、JMIM[13] 和 DWUR[14] 等)。然而这些算法存在忽视对交互依赖特征相关性和冗余性判断的问题。

因此，本文提出一种利用联合互信息和互信息判断特征与类标签之间相关性和冗余性的特征选择算法 (joint feature relevance and redundancy, JFRR)。该算法利用联合互信息计算在已选特征下候选特征与类标签之间的相关性；通过互信息计算已选特征和候选特征的冗余性；通过在 9 个基准基因数据集的实验对比，该算法 (JFRR) 优于其他特征选择算法 (MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR[15])。

1 联合互信息的一些概念

设X、Y和Z是 3 个离散型变量[16]，其中，$X=$

收稿日期：2021－05－18；修回日期：2022－04－28
基金项目：国家科技基础性工作专项 (2015FY111700-6)
作者简介：张俐 (1977－)，男，博士，副教授，主要从事特征工程与机器学习等方面的研究.
*通信作者：张俐，E-mail：zhangli_3913@163.com

$\{x_1,x_2,\cdots,x_L\}$，$Y=\{y_1,y_2,\cdots,y_M\}$，$Z=\{z_1,z_2,\cdots,z_N\}$。因此，X和Y之间的互信息定义如下：

$$I(X;Y)=\sum_{i=1}^{L}\sum_{j=1}^{M}p(x_i,y_i)\log_2\frac{p(x_i,y_i)}{p(x_i)p(y_i)} \tag{1}$$

式中，$p(x_i,y_i)$指联合分布；$p(x_i)$和$p(y_j)$指边际分布。

同时，X、Y和Z的条件互信息定义如下：

$$I(X;Y|Z)=\sum_{t=1}^{N}p(z_t)\sum_{i=1}^{L}\sum_{j=1}^{M}p(x_i,y_i|z_t)\times \log_2\frac{p(x_i,y_i|z_t)}{p(x_i|z_t)p(y_i|z_t)} \tag{2}$$

根据文献 [13] 的定义，联合互信息定义如下：

$$I(X,Y;Z)=I(Y;Z)+I(X;Z|Y) \tag{3}$$

2　JFRR 算法的提出

通过以上描述可知，传统的特征选择算法通常使用最小化冗余项和最大化相关项选择特征子集S。但是由此产生如下问题：1) 当已选特征量增加时，冗余项的大小也会随着相关项的增加而增加。这就存在一些冗余特征可能被选中；2) 在冗余项中，只考虑已选特征和候选特征之间互信息的计算，而忽视类标签，可能会造成已选特征和候选特征共享信息，意味着它们之间存在冗余信息。事实上，它们可能与类标签集合C之间共享不同信息。

以上问题可能会高估某些候选特征的重要性[17-19]。因此需要考虑，如何在已选特征集合S规模不断增加的情况下，解决S与类标签集合C的相关性，同时解决候选特征f_k与S的冗余性，以及解决在S条件下，候选特征f_k与类标签C的相关性的问题。

为此，本文提出一种基于信息论的特征选择算法 (JFRR)。该算法充分利用了线性累计加和的方式，具体如下：

$$J_{\mathrm{JFRR}}=\sum_{f_i\in S}(I(f_k,f_i;C)-I(f_k;f_i)) \tag{4}$$

式中，设F是原始特征集合，$S\subset F$；$J(\cdot)$代表评估标准；$f_i\in S,f_k\in F-S$。

通过式 (4) 可知，JFRR 算法利用联合互信息和互信息原理充分考虑S与C之间的相关性，f_k与S的冗余性以及在S条件下，f_k与C之间的相关性。JFRR 算法的具体描述如下。

输入：原始特征集合$F=\{f_1,f_2,\cdots f_n\}$，类标签集合C，已选特征子集S，阈值K

输出：最优特征子集S

1) 初始化：$S=\varnothing$, $k=0$

2) for k=1 to n

3) 计算每个特征与标签的互信息值$I(C;f_k)$

4) End for

5) $J_{\mathrm{JFRR}}(f_k)=\arg\max(I(f_k;C))$

6) Set $F\leftarrow F\backslash\{f_k\}$

7) Set $S\leftarrow\{f_k\}$

8) while $k\leqslant K$

9) for each $f_k\in F$ do

10) 根据式 (1)，计算f_k与f_i之间冗余$I(f_k;f_i)$的值；

11) 根据式 (1)，计算f_i与C之间相关性$I(f_i;C)$的值；

12) 根据式 (3)，计算f_k、f_i与C之间联合互信息$I(f_k,f_i;C)$的值；

13) 根据式 (4)，更新$J_{\mathrm{JFRR}}(f_k)$的值；

14) end for

15) 根据$J_{\mathrm{JFRR}}(f_k)$评估标准，寻找最优的候选特征f_k

16) Set $F\leftarrow F\backslash\{f_k\}$

17) Set $S\leftarrow\{f_k\}$

18) k=k+1

19) end while

从式 (4) 可知，JFRR 算法采用前向顺序搜索特征子集。JFRR 算法主要分为 3 部分。第 1 部分为 1)～7)，主要是初始化S集合和计数器k；将选择出最大的特征f_k加入S集合，同时f_k变成已选特征f_i。第 2 部分为 8)～13)，分别计算$I(f_i;C)$、$I(f_k;f_i)$和$I(f_k,f_i;C)$的值。第 3 部分为 14)～19)，根据式 (4) 的选择标准选择f_k，一直循环到用户指定的阈值K就停止循环。

3　实验验证与分析

3.1　实验环境设置

本节将 JFRR 与 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 算法进行对比。具体分类器为：决策树 (C4.5) 和支持向量机 (support vector machine, SVM)。本文的实验环境是 Intel-i7 处理器，16 GB 内存，仿真软件是 Python2.7。实验数据集选择 ASU 和 UCI 基因数据集[9, 20]，详细描述见表 1。其中，这 9 个数据集包含不同的样本数、特征数和类数。样本范围为 50～569，特征范围为 31～9 712，类的范围为 2～12，数据类型涉及连续

型和离散型。采用 6 折交叉验证方法进行实验验证。为保证实验公平，分别通过分类评价指标 fmc (F1_micro) 和 pcm(Precision_micro) 来评价预测性能。

表 1 数据集描述

序号	数据集	样本数	特征数	分类标签数	数据来源
1	lung	203	3 312	5	ASU
2	lung_discrete	73	325	7	ASU
3	lymphoma	96	4 026	9	ASU
4	Carcinom	174	9 182	11	ASU
5	nci9	60	9 712	9	ASU
6	GLIOMA	50	4 434	4	ASU
7	dermatology	358	35	6	UCI
8	wdbc	569	31	2	UCI
9	arrhythmia	416	279	12	UCI

3.2 特征选择算法性能比较与讨论

为了比较 JFRR 与 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 算法之间的优劣性，将它们所选的特征子集放到同一个分类器 (C4.5 和 SVM) 进行比较，特征子集的规模设置为 30。表 2 选择 C4.5 分类器。表 3 选择 SVM 分类器。在表 2～表 3 中，粗体代表该数据集下特征选择算法中最高平均分类预测值。“Wins/Ties/Losses”描述 JFRR 算法分别与 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 算法之间的优/平/输个数。

3.2.1 特征选择算法的 fmc 性能比较

在表 2 中，7 个特征选择算法的平均 fmc 精度值分别为 82.459%、80.24%、68.122%、75.356%、68.695%、73.047% 和 77.296%。JFRR 算法获得最高 fmc 值。同时，从 WINS/TIES/LOSSES 行的统计结果得出 JFRR 分别优于 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 算法 9、9、9、9、8 和 6 次。

表 2 C4.5 分类器的平均 fmc 性能比较

%

数据集	JFRR	MID	MIQ	CMIM	JMIM	CFR	CMI-MRMR
lung	87.555	86.546	79.907	84.632	81.804	75.266	**90.526**
lung_discrete	**86.039**	80.722	63.346	77.378	63.531	77.718	84.493
lymphoma	**89.322**	86.672	67.11	86.794	61.462	85.761	87.835
Carcinom	**77.595**	72.123	58.313	53.895	58.294	50.884	64.932
nci9	**75.554**	72.605	48.581	72.903	55.971	69.034	46.108
GLIOMA	**80.011**	79.448	58.68	61.055	58.496	54.448	74.627
dermatology	**94.41**	93.017	93.298	93.341	94.175	93.572	**94.41**
wdbc	**95.966**	95.789	94.738	95.445	94.557	94.38	95.259
arrhythmia	55.677	55.235	49.122	52.762	49.962	56.36	**57.473**
平均值	**82.459**	80.24	68.122	75.356	68.695	73.047	77.296
WINS/TIES/LOSSES		9/0/0	9/0/0	9/0/0	9/0/0	8/0/1	6/1/2

表 3 SVM 分类器的平均 fmc 性能比较

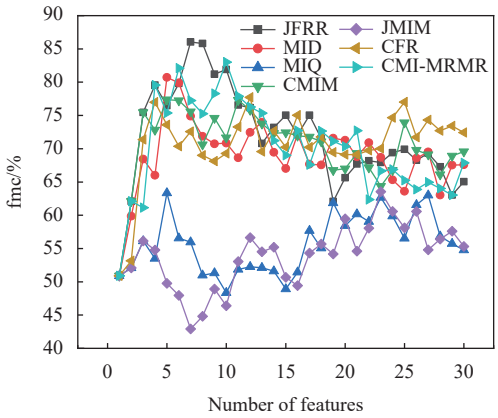
%

数据集	JFRR	MID	MIQ	CMIM	JMIM	CFR	CMI-MRMR
lung	91.106	90.111	77.344	89.126	84.694	85.184	**92.563**
lung_discrete	**91.906**	87.767	66.539	86.272	61.304	83.985	87.49
lymphoma	**95.102**	95.102	70.741	93.713	65.171	91.959	95.102
Carcinom	88.653	**89.693**	74.39	74.702	74.39	62.107	87.826
nci9	**83.15**	79.304	48.168	76.868	53.494	75.737	48.168
GLIOMA	32.165	32.165	32.165	34.248	30.081	34.248	**36.331**
dermatology	92.432	91.876	91.876	91.867	**97.466**	92.432	91.876
wdbc	**94.91**	94.563	90.677	94.38	90.852	94.559	90.333
arrhythmia	**59.445**	58.746	57.509	57.509	57.509	57.509	58.464
平均值	**80.985**	79.925	67.712	77.631 7	68.329	75.302	76.461
WINS/TIES/LOSSES		6/2/1	8/1/0	8/0/1	8/0/1	7/1/1	6/1/2

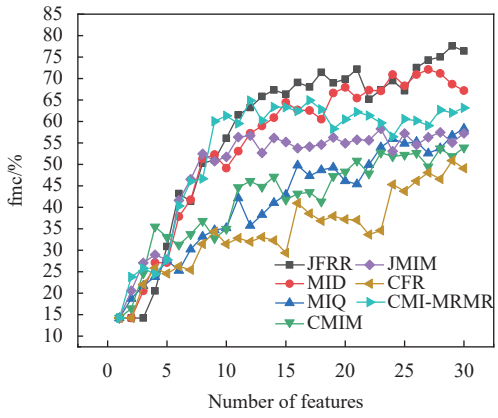
在表 3 中，7 个特征选择算法的平均 fmc 精度值分别为 80.985%、79.925%、67.712%、77.631 7%、68.329%、75.302% 和 76.461%。JFRR 算法获得最高 fmc 值。同时，从 WINS/TIES/LOSSES 行的统

计结果得出 JFRR 分别优于 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 算法 6、8、8、8、7 和 6 次。

为了进一步比较特征子集对 fmc 值的影响，图 1 和图 2 分别给出部分数据集的 fmc 性能差异。当数据的维数不断增加时，JFRR 算法通过动态调整特征间的相关性和冗余性提升了特征子集的数据质量。图 1 和图 2 的实验结果显示，JFRR 算法对分类提升的效果明显。并且，JFRR 明显优于 MID、CMIM、MIQ、JMIM、CFR 和 CMI-MRMR。



a. lung_discrete

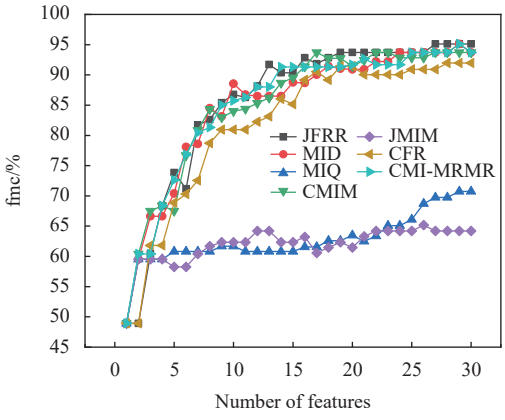


b. Carcinom

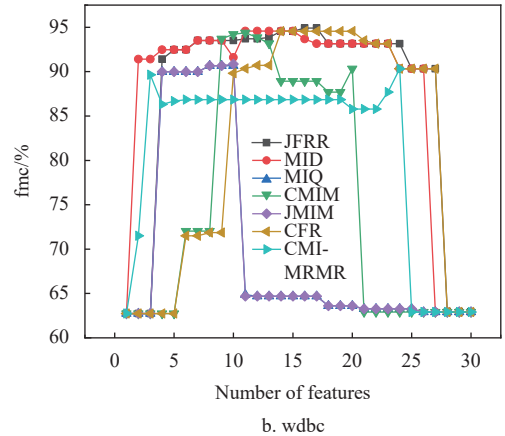
图 1　C4.5 在高维数据集上的性能比较

图 1 是 C4.5 在高维数据集上的性能比较。在图 1a 中，JFRR 算法的分类 fmc 值为 86.039%，是 7 种分类算法中最高的，分别比 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 高出 5.317%、22.693%、8.661%、22.508%、8.321% 和 1.546%。在图 1b 中，JFRR 算法的分类 fmc 值为 77.595%，也是 7 种分类算法中最高的，分别比 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 高出 5.472%、19.282%、23.7%、19.301%、26.711% 和 12.663%。

图 2 是 SVM 在高维数据集上的性能比较。在图 2a 中，JFRR 算法的分类 fmc 值为 95.102%，是 7 种分类算法中最高的，分别比 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 高出 0.0%、24.361%、1.389%、29.931% 和 3.143% 和 0.0%。在图 2b 中，JFRR 算法的分类 fmc 值为 94.91%，是 7 种分类算法中最高的，分别比 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 高出 0.347%、4.233%、0.53%、4.058%、0.351% 和 4.577%。



a. lymphoma

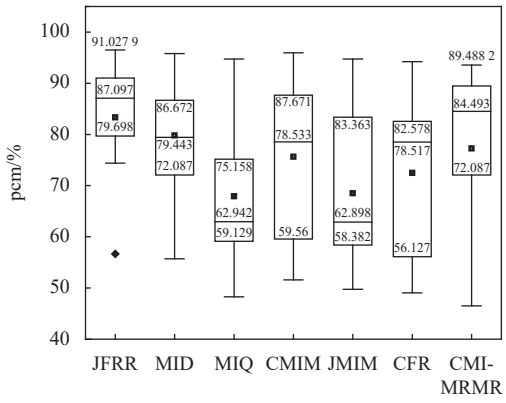


b. wdbc

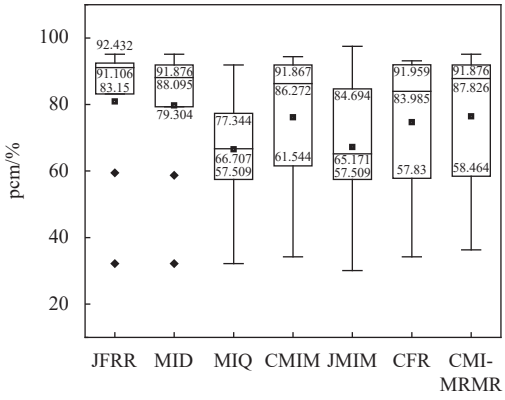
图 2　SVM 在高维数据集上的性能比较

3.2.2　特征选择算法的 pcm 性能比较

图 3 为 pcm 盒图。从图 3a 中可以得出，在 C4.5 分类器的 pcm 盒图中，使用 JFRR 算法选择出的特征集合在五位数 (最小值、四分位数 (第 25 个百分位数)、中位数、四分位数 (第 75 个百分位数) 和最大值) 中体现出的分类效果都是最优。同时，从图 3b 中也可以得出，在 SVM 分类器的 pcm 盒图中，使用 JFRR 算法选择出的特征集合在五位数 (最小值、四分位数 (第 25 个百分位数)、中位数和四分位数 (第 75 个百分位数)) 中体现出的分类效果都是最优的效果。



a. C4.5



b. SVM

图 3 C4.5 分类器和 SVM 分类器的 pcm 盒图

综上，不同分类器表现出的分类结果不尽相同。但是，JFRR 算法在 fmc 和 pcm 的评价指标值在大多数数据集上都是最好。从 C4.5 和 SVM 分类器表现结果中可知，C4.5 分类性能明显优于 SVM 分类性能。

3.3 算法的运行时间分析

计算特征选择算法的运行时间也是衡量特征选择算法重要性的标准之一。JFRR、MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 算法在 9 个数据集上进行特征排序后得出的运行时间如表 4 所示。可以看出，JFRR 算法的运行时间在可接受的范围之内。

3.4 实验算法比较

本节分析 JFRR 与 MID、CMIM、MIQ、JMIM、CFR 和 CMI-MRMR 之间在交互特征依赖相关性和冗余性的差异。从表 5 可以得出，与 JFRR 相比，MID、MIQ、CMIM 和 CFR 将$I(f_k;C)$定义为衡量特征相关性的标准。CMI-MRMR 将$I(f_i,C|f_k)$定义为衡量特征相关性的标准。只有 JFRR 和 JMIM 将$I(f_k,f_i;C)$定义为衡量交互特征依赖性动态变化标准。但是，JMIM 算法却忽视特征冗余性变化。因此，得出 JFRR 与其他特征选择算法差异明显。

表 4 不同特征选择算法运行时间比较

s

数据集	JFRR	MID	MIQ	CMIM	JMIM	CFR	CMI-MRMR
lung	95.568	118.655	57.357	127.739	109.277	126.454	882.251
lung_discrete	2.718	0.969	1.0	2.796	2.721	2.781	31.682
lymphoma	37.508	9.497	9.763	27.825	27.308	27.966	326.731
Carcinom	198.758	88.56	100.252	212.276	369.935	369.935	2 298.744
nci9	76.264	27.323	25.375	51.558	50.038	48.722	25.375
GLIOMA	46.543	20.548	22.712	38.615	70.993	65.942	353.598
dermatology	0.868	0.31	0.318	0.55	0.551	0.554	6.811
wdbc	1.434	0.598	0.591	1.31	1.311	1.285	14.378
arrhythmia	15.985	5.952	8.217	19.022	23.048	16.727	214.663
平均值	52.85	30.268	25.065	53.521	72.798	73.374	461.581

表 5 算法比较

算法	考虑特征之间的交互相关性变化	特征冗余性
MID	$I(f_k;C)$	是
CMIM	$I(f_k;C)$	是
MIQ	$I(f_k;C)$	是
JMIM	$I(f_k,f_i;C)$	否
CFR	$I(f_k;C)$	是
JFRR	$I(f_k,f_i;C)$	是
CMI-MRMR	$I(f_i,C\|f_k)$	是

4 结 束 语

随着基因数据中高维特征数据的不断增多，特征间的关系变得越来越复杂 (包含大量无关特征和冗余特征)。而传统的特征选择算法往往忽视特征间的相关性和冗余性之间的联系。本文提出一种基于联合互信息的 JFRR 算法。该算法利用互信息和联合互信息间的关系动态分析和调整特征间以及特征与类标签间的相关信息和冗余信息，从而达到删除无关特征和冗余特征的目的，以此提高特征子集的数据质量。为了全面验证 JFRR 算法的有效性，实验在 9 个基因数据集上进行。分别通过使用分类器 (C4.5 和 SVM) 和分类准确率指标 (fmc 和 pcm) 全面评估所选特征子集的质量。实验结果证明

JFRR 明显优于 MID、MIQ、CMIM、JMIM、CFR 和 CMI-MRMR 等 6 种特征选择算法。

但在一些基因数据中，JFRR 算法仍旧存在选择出的特征子集不理想的情况。未来的工作将进一步研究和改进互信息和联合互信息的关系，并以此优化 JFRR 算法，同时在更广泛的基因数据集中对算法进行验证，以此提高分类预测精度。

参 考 文 献

[1] DABBA A, ABDELKAMEL T, SAMY M, et al. Gene selection and classification of microarray data method based on mutual information and moth flame algorithm[J]. Expert Systems with Applications, 2021, 166: 114012.

[2] HAMBALI M A, OLADELE T O, ADEWOLE K S. Microarray cancer feature selection: Review, challenges and research directions[J]. International Journal of Cognitive Computing in Engineering, 2020, 1: 78-97.

[3] 王翔, 胡学钢. 高维小样本分类问题中特征选择研究综述[J]. 计算机应用, 2017, 37(9): 2433-2438.
WANG X, HU X G. Overview on feature selection in high-dimensional and small-sample-size classification[J]. Journal of Computer Applications, 2017, 37(9): 2433-2438.

[4] WANG X, LIU J, CHENG Y, et al. Dual hypergraph regularized PCA for biclustering of tumor gene expression data[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(12): 2292-2303.

[5] LIU H, GREGORY D. A semi-parallel framework for greedy information-theoretic feature selection[J]. Information Sciences, 2019, 492: 13-28.

[6] CAI J, LUO J W, WANG S L, et al. Feature selection in machine learning: A new perspective[J]. Neurocomputing, 2018, 300: 70-79.

[7] LEE C Y, CAI J Y. LASSO variable selection in data envelopment analysis with small datasets[J]. Omega, 2020, 91: 102019.

[8] GAO L Y, WU W G. Relevance assignation feature selection method based on mutual information for machine learning[J]. Knowledge-Based Systems, 2020, 209: 106439.

[9] 谢娟英, 王明钊, 周颖, 等. 非平衡基因数据的差异表达基因选择算法研究[J]. 计算机学报, 2019, 42(6): 1232-1251.
XIE J Y, WANG M Z, ZHOU Y, et al. Differential expression gene selection algorithms for unbalanced gene datasets[J]. Chinese Journal of Computers, 2019, 42(6): 1232-1251.

[10] MACEDO F, OLIVEIRA M R, PACHECO A, et al. Theoretical foundations of forward feature selection methods based on mutual information[J]. Neurocomputing, 2019, 325: 67-89.

[11] GAO W F, HU L, ZHANG P, et al. Feature selection considering the composition of feature relevancy[J]. Pattern Recognition Letters, 2018, 112: 70-74.

[12] BROWN G, POCOCK A, ZHAO M J, et al. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection[J]. The Journal of Machine Learning Research, 2012, 13: 27-66.

[13] BENNASAR M, HICKS Y, SETCHI R. Feature selection using joint mutual information maximisation[J]. Expert Systems with Applications, 2015, 42(22): 8520-8532.

[14] 肖利军, 郭继昌, 顾翔元. 一种采用冗余性动态权重的特征选择算法[J]. 西安电子科技大学学报, 2019, 46(5): 155-161.
XIAO L J, GUO J C, GU X Y. Algorithm for selection of features based on dynamic weights using redundancy[J]. Journal of XiDian University. 2019, 46(5): 155-161.

[15] GU X Y, GUO J C, XIAO L J, et al. Conditional mutual information-based feature selection algorithm for maximal relevance minimal redundancy[J]. Applied Intelligence, 2022, 52(2): 1436-1447.

[16] MEYER P E, SCHRETTER C, BONTEMPI G. Information-Theoretic feature selection in microarray data using variable complementarity[J]. IEEE Journal of Selected Topics in Signal Processing, 2008, 2(3): 261-274.

[17] ZHANG P, GAO W F. Feature selection considering uncertainty change ratio of the class label[J]. Applied Soft Computing, 2020, 95: 106537.

[18] CHE J X, YANG Y L, LI L, et al. Maximum relevance minimum common redundancy feature selection for nonlinear data[J]. Information Sciences, 2017, 409-410: 68-86.

[19] ZHANG Y S, ZHANG Q, CHEN Z J, et al. Feature assessment and ranking for classification with nonlinear sparse representation and approximate dependence analysis[J]. Decision Support Systems, 2019, 122: 113064.

[20] 谢娟英, 丁丽娟, 王明钊. 基于谱聚类的无监督特征选择算法[J]. 软件学报, 2020, 31(4): 1009-1024.
XIE J Y, DING L J, WANG M Z. Spectral clustering based unsupervised feature selection algorithms[J]. Journal of Software, 2020, 31(4): 1009-1024.

编　辑　蒋　晓