软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, 2022,33(4):1338−1353 [doi: 10.13328/j.cnki.jos.006466]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 基于可辨识矩阵的完全自适应 2D 特征选择算法*

谢娟英, 吴肇中

(陕西师范大学 计算机科学学院, 陕西 西安 710119)

通信作者: 谢娟英, E-mail: xiejuany@snnu.edu.cn

**摘 要**: 针对基于信息增益与皮尔森相关系数的特征选择算法 FSIP (feature selection based on information gain and Pearson correlation coefficient)存在的特征子集选取需要人工参与的问题, 提出基于可辨识矩阵的完全自适应 2D 特征选择算法 DFSIP (discernibility based FSIP). DFSIP 算法完全自适应地发现特征子集, 每次选择当前特征中最重要的一个特征, 并以此特征约简可辨识矩阵, 剔除冗余特征, 最终自适应地获得最优特征子集. 依据最优特征子集构建 *K*-ELM 分类器来评价最优特征子集的类别辨识能力. 在基因数据集的实验测试以及与 FSIP, mRMR, LLE Score, DRJMIM, AVC, AMID 算法的实验比较和统计重要性检测表明: DFSIP 算法能够自动选择出辨识能力更强的特征子集, 基于此特征子集的分类器具有很好的分类性能.

**关键词**: 可辨识矩阵; 特征辨识度; 特征独立性; 特征选择; 信息增益; 皮尔森相关系数

**中图法分类号**: TP18

中文引用格式: 谢娟英, 吴肇中. 基于可辨识矩阵的完全自适应 2D 特征选择算法. 软件学报, 2022, 33(4): 1338−1353. http://www.jos.org.cn/1000-9825/6466.htm

英文引用格式: Xie JY, Wu ZZ. Totally Adaptive 2D Feature Selection Algorithm Based on Discernibility Matrix. Ruan Jian Xue Bao/ Journal of Software, 2022, 33(4): 1338−1353 (in Chinese). http://www.jos.org.cn/1000-9825/6466.htm

## Totally Adaptive 2D Feature Selection Algorithm Based on Discernibility Matrix

XIE Juan-Ying, WU Zhao-Zhong

(School of Computer Science, Shaanxi Normal University, Xi'an 710119, China)

**Abstract**: To overcome the limitations of the FSIP (feature selection based on information gain and Pearson correlation coefficient) feature selection algorithm that need human to determine the borderline to detect the feature subsets, the totally adaptive 2D feature selection algorithm is proposed in this study based on discernibility matrix. It is referred to as DFSIP (discernibility based FSIP). DFSIP introduces discernibility matrix into the feature selection process of FSIP. It first initializes the candidate feature set comprising all features and constructs the initial discernibility matrix, then it detects the most significant feature from the current candidate feature set, so as to add it to feature subset and use it to reduce the discernibility matrix. After that the candidate feature set is updated using the union of the cells of the reduced discernibility matrix, and the most significant feature is detected from the current candidate feature set again, so as to put it into the feature subset and use it to reduce the discernibility matrix, and the candidate feature set is updated again. This process repeats till there is not any feature left in the candidate feature set. The power of DFSIP is tested on very famous gene expression datasets, and its performance is compared with that of the popular feature selection algorithms including FSIP, mRMR, LLE Score, DRJMIM, AVC, and AMID by comparing the performance of the *K*-ELM classifier built using the feature subset detected by these feature selection algorithms. In addition, the significant test is done to verify whether or not there is the significant difference between DFSIP and FSIP as well as other compared feature selection algorithms. The experimental results demonstrate that DFSIP is superior to the compared ones, especially it has the significant difference to LLE Score, DRJMIM, and AMID feature selection algorithms. Although there is not

* 基金项目: 国家自然科学基金(62076159, 61673251, 12031010); 国家重点研发计划(2016YFC0901900); 中央高校基本科研业务费专项资金(GK202105003); 研究生培养创新基金(2016CSY009, 2018TS078)

本文由"面向开放场景的鲁棒机器学习"专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.

收稿时间: 2021-03-10; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2021-10-26

significant difference between DFSIP and FSIP, it defeats FSIP in performance. It can be concluded that DFSIP can totally adaptively detect the feature subset with sound classification capability.

**Key words**: discernibility matrix; feature discernibility; feature independence; feature selection; information gain; Pearson correlation coefficient

科学与技术的发展, 带来跨学科领域的研究与日俱增, 基于人工智能技术的生物医学大数据分析得到机器学习、数据挖掘等人工智能领域学者的关注[1−3]. 然而, 生物医学数据往往具有高维小样本特点, 特征数远多于样本数, 引发维数灾难, 带来大量冗余或无关特征. 剔除冗余和无关特征是分析该类生物医学大数据的基础和首要步骤, 使得特征选择成为当前研究热点之一[4−8].

特征选择研究依据搜索策略可分为基于全局最优搜索策略的特征选择方法、采用随机搜索策略的特征选择方法、采用启发式搜索策略的特征选择方法, 依据与分类器的关系则可分为 Filter 特征选择方法、Wrapper 特征选择方法和 Embedded 特征选择方法等[9]. Filter 特征选择方法不依赖于具体的学习机, 依据给定的评价准则选择相应的特征构成特征子集, 速度快, 但是需要事先给定阈值作为停止准则. 距离度量[10−12]、一致性度量[13−17]、相关性度量[18−21]和信息度量[22−32]是 Filter 特征选择算法的常用评价准则[33], 如 Laplacian 得分[34]、Constraint 得分[20]、Fisher 得分[35]、Pearson 相关系数[36]、互信息[23]、MIC[32]等. Wrapper 方法依赖于具体的学习机, 以学习机的分类性能评价特征子集的分类能力, 需要将训练集分为训练子集和验证子集, 非常费时, 且存在过适应风险. Embedded 方法也依赖于学习机, 但是与 Wrapper 方法不同, Embedded 方法不需要将训练集划分为训练子集和验证子集, 特征选择在优化学习机目标函数的过程中实现, 其缺点是设计优化目标函数非常困难. Filter 方法由于快速、不存在过适应而得到广泛应用和研究.

针对 Filter 方法需要给定阈值的缺陷, FSIP (Feature Selection based on Information gain and Pearson correlation coefficient)算法[37]提出了基于特征辨识度与独立性的 2D 可视化特征选择思想, 以信息增益定义特征辨识度, 以 Pearson 相关系数定义特征的独立性, 构造以辨识度和独立性分别作为横、纵坐标的 2D 空间, 所有特征被展示在该 2D 空间, 使得辨识度和独立性都很强的特征位于空间右上角区域, 远离右下角区域的特征. 为了量化特征对于分类的贡献, 定义特征的重要度为其辨识度与独立性之积, 及其坐标确定的矩形面积, 选择对分类贡献远大于其余特征的特征构成特征子集. 但是 FSIP 算法需要人为观测特征的 2D 空间分布, 实现特征选择, 没有实现特征选择的完全自动化. 为此, 本文提出 DFSIP (discernibility based FSIP)算法, 以期完全自适应地发现特征子集, 实现完全自动化的特征选择. DFSIP 算法引入可辨识矩阵, 选取当前最优特征, 用当前最优特征约简可辨识矩阵, 约简后的可辨识矩阵的非空元素之并集构成新候选特征, 从候选特征中选择最优特征, 以该最优特征再次约简可辨识矩阵, 新约简后的可辨识矩阵的非空元素再构成候选特征, 再选择当前候选特征中的最优特征. 反复迭代, 直至可辨识矩阵的每个元素为空, 也即候选特征集合为空集停止. 此时, 被选择的最优特征构成特征子集. DFSIP 使 FSIP 算法的人工参与选择特征子集的过程升级为完全自动地选择特征子集的过程, 实现了特征子集的完全自适应发现.

## 1 信息熵

### 1.1 信息熵

熵(entropy)是 1877 年物理学家玻尔兹曼[38]提出的一种状态函数, 被用于表示系统的状态, 系统越无序, 其熵越大. 后被使用到信息论领域, 用于表示系统的信息含量, 即系统越有序、确定, 其信息熵值越小.

假设一个信息系统的变量有 $m$ 个, 表示为集合 $\boldsymbol{U}=\{u_1,u_2,\ldots,u_m\}$, $p(u_i)$是变量 $u_i$ 的概率, 则该系统的 $m$ 个变量可视为一个随机变量的 $m$ 种取值, 那么该系统的信息熵可以表达为集合 $\boldsymbol{U}$ 的信息熵 $H(\boldsymbol{U})$, 定义为公式(1):

$$H(\boldsymbol{U})=-\sum_{u}p(u)\log p(u)=-\sum_{i=1}^{m}p(u_i)\log p(u_i) \tag{1}$$

$H(\boldsymbol{U})$的值越大, 代表该系统越不稳定, 不确定性越大, 包含信息量越多.

### 1.2 联合熵

联合熵(joint entropy), 表示两个变量集合同时考虑时的信息熵. 对于给定变量集合 $\boldsymbol{U}$ 和包含 $n$ 个变量的集合 $\boldsymbol{V}=\{v_1,v_2,\ldots,v_n\}$, 其联合熵 $H(\boldsymbol{U},\boldsymbol{V})$定义为公式(2):

$$H(\boldsymbol{U},\boldsymbol{V})=-\sum_{u,v}p(u,v)\log p(u,v)=-\sum_{j=1}^{n}\sum_{i=m}^{m}p(u_i,v_j)\log p(u_i,v_j) \tag{2}$$

其中, $p(u,v)$为变量 $u$ 和 $v$ 的联合概率.

### 1.3 条件熵

条件熵(conditional entropy), 表示某变量集合在另一变量集合确定条件下的信息熵. 对于给定变量集合 $\boldsymbol{U}$ 和给定变量集合 $\boldsymbol{V}$, 其条件熵 $H(\boldsymbol{U}|\boldsymbol{V})$定义为公式(3):

$$H(\boldsymbol{U}\mid\boldsymbol{V})=\sum_{j=1}^{n}p(v_j)H(\boldsymbol{U}\mid V=v_j)=-\sum_{j=1}^{n}p(v_j)\sum_{i=1}^{m}p(u_i\mid v_j)\log p(u_i\mid v_j)=-\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i\mid v_j) \tag{3}$$

其中, $p(u|v)$为变量 $u$ 在 $v$ 给定条件下的条件概率.

### 1.4 信息增益

信息增益(information gain)定义为公式(4), 表示集合 $\boldsymbol{U}$ 的信息量与其在集合 $\boldsymbol{V}$ 条件下信息量的差值:

$$IG(\boldsymbol{U},\boldsymbol{V})=H(\boldsymbol{U})-H(\boldsymbol{U}|\boldsymbol{V}) \tag{4}$$

当信息增益用于特征选择时, 表示类标熵 $H(\boldsymbol{U})$在特征 $\boldsymbol{V}$ 给定条件下的信息量减少程度, 度量了类标信息量在给定特征 $\boldsymbol{V}$ 条件下更加确定了多少, 表达了特征 $\boldsymbol{V}$ 对分类的贡献.

条件概率公式如公式(5)所示:

$$p(u\mid v)=\frac{p(u,v)}{p(v)} \tag{5}$$

全概率公式如公式(6)所示:

$$p(u)=\sum_{i=1}^{n}p(v_i)\,p(u\mid v_i) \tag{6}$$

将公式(1)–公式(3)、公式(5)、公式(6)带入公式(4)中, 可得公式(7), 表达了信息增益、信息熵、联合熵、条件熵的关系:

$$\begin{aligned}
IG(\boldsymbol{U},\boldsymbol{V})&=H(\boldsymbol{U})-H(\boldsymbol{U}\mid\boldsymbol{V})\\
&=-\sum_{i=1}^{m}p(u_i)\log p(u_i)-\left(-\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i\mid v_j)\right)\\
&=-\sum_{i=1}^{m}p(u_i)\log p(u_i)-\left(-\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log\frac{p(u_i,v_j)}{p(v_j)}\right)\\
&=-\sum_{i=1}^{m}p(u_i)\log p(u_i)-\left(-\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i,v_j)+\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(v_j)\right)\\
&=-\sum_{i=1}^{m}p(u_i)\log p(u_i)-\left(-\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i,v_j)+\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i)\,p(v_j\mid u_i)\log p(v_j)\right)\\
&=-\sum_{i=1}^{m}p(u_i)\log p(u_i)-\left(-\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i,v_j)+\sum_{j=1}^{n}p(v_j)\log p(v_j)\right)\\
&=H(\boldsymbol{U})-(H(\boldsymbol{U},\boldsymbol{V})-H(\boldsymbol{V}))\\
&=H(\boldsymbol{U})+H(\boldsymbol{V})-H(\boldsymbol{U},\boldsymbol{V})
\end{aligned} \tag{7}$$

对公式(7)进一步推导, 可得公式(8):

$$\begin{aligned}IG(\boldsymbol{U},\boldsymbol{V}) &= H(\boldsymbol{U})+H(\boldsymbol{V})-H(\boldsymbol{U},\boldsymbol{V})\\ &= -\sum_{i=1}^{m}p(u_i)\log p(u_i)-\sum_{j=1}^{n}p(v_j)\log p(v_j)+\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i,v_j)\\ &= -\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i)-\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(v_j)+\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log p(u_i,v_j)\\ &= -\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)(\log p(u_i)+\log p(v_j)-\log p(u_i,v_j))\\ &= -\sum_{j=1}^{n}\sum_{i=1}^{m}p(u_i,v_j)\log\frac{p(u_i)p(v_j)}{p(u_i,v_j)}\end{aligned} \tag{8}$$

## 2 可辨识矩阵

可辨识矩阵表达了样本之间的可区分属性, 源于经典粗糙集[39], 在经典粗糙集发展为邻域粗糙集后, 被引入到邻域粗糙集.

### 2.1 经典粗糙集

经典粗糙集模型由 Pawlak 教授[39]于 1982 年提出, 以四元组 $\boldsymbol{S}=\langle \boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$表示一个信息系统, 其中: $\boldsymbol{X}$ 是对象集; $\boldsymbol{A}$ 是属性集, 也就是变量或者称为特征构成的集合; $\boldsymbol{V}=\bigcup \boldsymbol{V}_a$ , $\boldsymbol{V}_a$ 是特征 $a\in\boldsymbol{A}$ 的值域; $\rho$: $\boldsymbol{X}\times\boldsymbol{A}\to\boldsymbol{V}$ 是一个信息函数, $\rho_x$: $\boldsymbol{A}\to\boldsymbol{V}$. 令 $\boldsymbol{A}=\boldsymbol{C}\cup\boldsymbol{D}$, $\boldsymbol{C}$ 表示条件属性集(特征集), $\boldsymbol{D}$ 表示决策属性集(类别集), 本文默认 $\boldsymbol{D}$ 中仅包含一个元素, 即只有一个类标属性. $\forall \boldsymbol{E}\subseteq\boldsymbol{C}$, 根据 $\boldsymbol{E}$ 在 $\boldsymbol{X}$ 的属性取值得到 $\boldsymbol{X}$ 关于 $\boldsymbol{E}$ 的划分, 获得对象 $x_i\in\boldsymbol{X}$ 在属性集 $\boldsymbol{E}$ 上具有相同取值的对象集$[x_i]_{\boldsymbol{E}}$.

### 2.2 邻域粗糙集

经典粗糙集适于处理字符型数据, 不适于分析数值型数据. 为了克服这一局限, 2008 年, 胡等人[40]提出了邻域粗糙集, 定义邻域关系作为对经典粗糙集中等价关系的推广.

**定义 1(邻域关系[40])**. 四元组 $\boldsymbol{S}=\langle \boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$表示一个信息系统, 属性子集 $\boldsymbol{E}\subseteq\boldsymbol{A}$, 则该属性子集 $\boldsymbol{E}$ 下的邻域关系定义为公式(9):

$$\boldsymbol{N}_{\boldsymbol{E}}=\{(x_i,x_j)\in\boldsymbol{X}\times\boldsymbol{X}|\Delta^{\boldsymbol{E}}(x_i,x_j)\leqslant\delta,\ 0\leqslant\delta\leqslant 1\} \tag{9}$$

其中, $\Delta^{\boldsymbol{E}}$ 表示对象（样本）基于属性子集 $\boldsymbol{E}$ 的距离, 这里的距离可以是欧氏距离、$L_1$ 范数等任何距离函数, 本文后面的实验中都统一使用 $L_1$ 范数; $\delta$为阈值参数.

**定义 2(邻域[40])**. 已知四元组 $\boldsymbol{S}=\langle \boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$表示一个信息系统, 属性子集 $\boldsymbol{E}\subseteq\boldsymbol{A}$, $\boldsymbol{N}_{\boldsymbol{E}}$ 为对象集 $\boldsymbol{X}$ 关于 $\boldsymbol{E}$ 的邻域关系, 对于任意样本$\forall x_i\in\boldsymbol{X}$, 其在 $\boldsymbol{E}$ 下的邻域定义为公式(10):

$$\delta_{\boldsymbol{E}}(x_i)=\{x_j|x_j\in\boldsymbol{X},(x_i,x_j)\in\boldsymbol{N}_{\boldsymbol{E}}\} \tag{10}$$

**定义 3(邻域概率)**. 已知四元组 $\boldsymbol{S}=\langle \boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$表示一个信息系统, 属性子集 $\boldsymbol{E}\subseteq\boldsymbol{A}$, 对$\forall x_i\in\boldsymbol{X}$, 其邻域$\delta_{\boldsymbol{E}}(x_i)$在属性集 $\boldsymbol{E}$ 上的概率定义为公式(11), 即邻域内的样本数占总样本数的比例:

$$p(\delta_{\boldsymbol{E}}(x_i))=\frac{\|\delta_{\boldsymbol{E}}(x_i)\|}{\|\boldsymbol{X}\|} \tag{11}$$

其中, $\|\delta_{\boldsymbol{E}}(x_i)\|$, $\|\boldsymbol{X}\|$分别表示邻域$\delta_{\boldsymbol{E}}(x_i)$包含样本个数和信息系统总样本数.

**定义 4(邻域熵与邻域信息增益[41])**. 已知四元组 $\boldsymbol{S}=\langle \boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$表示一个信息系统, 属性子集 $\boldsymbol{E}\subseteq\boldsymbol{A}$, $\delta_{\boldsymbol{E}}(x_i)$, $\delta_{\boldsymbol{F}}(x_i)$分别为 $x_i$ 关于 $\boldsymbol{E}$, $\boldsymbol{F}$ 的邻域, 则邻域熵 $NH_\delta(\boldsymbol{E})$定义为公式(12):

$$NH_\delta(\boldsymbol{E})=-\sum_{x_i\in\boldsymbol{X}}p(\delta_{\boldsymbol{E}}(x_i))\log p(\delta_{\boldsymbol{E}}(x_i)) \tag{12}$$

由于对样本 $x_i$ 来说, 其在属性集 $\boldsymbol{E}$, $\boldsymbol{F}$ 上的邻域和邻域关系是唯一确定的, 故根据公式(8)的信息增益, 可定义邻域信息增益 $NIG_\delta(\boldsymbol{E},\boldsymbol{F})$为公式(13):

$$NIG_{\delta}(\boldsymbol{E},\boldsymbol{S})=-\sum_{x_i\in\boldsymbol{X}}p(\delta_{\boldsymbol{E}\cup\boldsymbol{S}}(x_i))\log\frac{p(\delta_{\boldsymbol{E}}(x_i))p(\delta_{\boldsymbol{S}}(x_i))}{p(\delta_{\boldsymbol{E}\cup\boldsymbol{S}}(x_i))} \tag{13}$$

### 2.3 可辨识矩阵

可辨识矩阵最早由 Skowron 于 1992 提出[42], 用以获得属性的约简, 即特征子集. 基于可辨识矩阵的特征选择研究是粗糙集特征选择研究的一大分支, 将其结合信息论的研究也有不少[41,43,44]. 假设已知一个信息系统的四元组表示 $\boldsymbol{S}=\langle\boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$, $\boldsymbol{A}=\boldsymbol{C}\cup\boldsymbol{D}$, $\boldsymbol{C}$ 表示条件属性集(特征集), $\boldsymbol{D}$ 表示决策属性集(类别集), 则基于邻域关系的可辨识矩阵及其约简等的概念可描述如下.

**定义 5(可辨识矩阵)**. 信息系统 $\boldsymbol{S}=\langle\boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$的可辨识矩阵 $\boldsymbol{M}$ 是一个$||\boldsymbol{X}||\times||\boldsymbol{X}||$的矩阵, 其元素 $\boldsymbol{m}_{ij}$ 定义为

$$\boldsymbol{m}_{ij}=\begin{cases}\{c\in\boldsymbol{C}\mid\Delta^{c}(x_i,x_j)\succ\delta,x_i\in\boldsymbol{X},x_j\in\boldsymbol{X}\},\ \boldsymbol{D}(x_i)\neq\boldsymbol{D}(x_j)\\ \varnothing, \qquad\qquad\qquad\qquad\qquad\text{otherwise}\end{cases} \tag{14}$$

其中, $\varnothing$表示空集, $\delta$是邻域关系定义公式(9)中的阈值参数, $\boldsymbol{D}(x_i)$是样本 $x_i$ 的类标, $\boldsymbol{m}_{ij}$ 是能区分样本 $x_i$, $x_j$ 的属性构成的集合.

**定义 6(约简)**. 一个信息系统 $\boldsymbol{S}=\langle\boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$的可辨识矩阵为 $\boldsymbol{M}$, 属性集 $\boldsymbol{E}\subseteq\boldsymbol{C}$, 则可辨识矩阵 $\boldsymbol{M}$ 关于 $\boldsymbol{E}$ 的约简定义为公式(15):

$$\boldsymbol{m}_{ij}=\begin{cases}\boldsymbol{m}_{ij},\ \boldsymbol{m}_{ij}\cap\boldsymbol{E}=\varnothing\\ \varnothing,\quad\text{otherwise}\end{cases} \tag{15}$$

其中, $\boldsymbol{m}_{ij}$ 为可辨识矩阵 $\boldsymbol{M}$ 的第$(i,j)$位置的元素, $\varnothing$表示空集.

**定义 7(候选特征子集 $\boldsymbol{CS}$)**. 信息系统 $\boldsymbol{S}=\langle\boldsymbol{X},\boldsymbol{A},\boldsymbol{V},\rho\rangle$的可辨识矩阵为 $\boldsymbol{M}$, $\exists\boldsymbol{E}\subset\boldsymbol{C}$, 则使用 $\boldsymbol{E}$ 对可辨识矩阵 $\boldsymbol{M}$ 进行约简, 得到新的可辨识矩阵 $\boldsymbol{M}'$, $\boldsymbol{M}'$中所有非空元素的并集, 即约简可辨识矩阵 $\boldsymbol{M}'$的非空元素所包含的所有特征构成了候选特征子集 $\boldsymbol{CS}$:

$$\boldsymbol{CS}=\{c\mid c\in\boldsymbol{C}\wedge c\in\cup\boldsymbol{m}_{ij},\boldsymbol{m}_{ij}\in\boldsymbol{M}',i=1,...,||\boldsymbol{X}||,j=1,...,||\boldsymbol{X}||\} \tag{16}$$

## 3 DFSIP 算法

DFSIP 算法将所有样本依据邻域关系分为自身邻域与非自身邻域, 并以此构建可辨识矩阵, 然后使用邻域信息增益[41]定义特征辨识度, 用皮尔逊相关系数定义特征独立性, 再以特征辨识度与独立性之积定义特征重要度, 迭代选择当前剩余特征里重要度值最大的特征, 并依据该选择的特征对可辨识矩阵进行约简, 直到可辨识矩阵为空, 即没有余下特征停止. 此时, 所有已经选择的特征构成特征子集. 该特征子集的特征数自动确定, 克服了 FSIP 需要人为参与的问题. 基于特征子集构造 K-ELM[45]分类器, 以分类器的分类性能评价特征子集的类别辨识能力, 从而评价本文 DFSIP 算法发现最优特征子集的能力.

### 3.1 特征辨识度、独立性与重要度

不妨设数据集 $\boldsymbol{Data}\in\Re^{n\times(m+1)}$, 即数据集包含 $n$ 个样本, 每个样本含 $m$ 个特征,$\boldsymbol{f}_i\in\Re^{n}$ 表示第 $i(i=1,\ldots,m)$个特征的特征向量,$\boldsymbol{Y}\in\Re^{n\times1}$ 为类标信息向量.

**定义 8(特征辨识度)**. 使用公式(13)定义的邻域信息增益来度量每个特征对于分类的贡献, 作为特征辨识度, 第 $i$ 个特征的辨识度 $dis_i$ 定义为公式(17), 表示引入第 $i$ 个特征带来的信息量变化:

$$dis_i=NIG_{\delta}(\boldsymbol{Y},\boldsymbol{f}_i) \tag{17}$$

此信息量变化越大, 说明第 $i$ 个特征对于分类的贡献越大.

**定义 9(特征独立性)**. 使用皮尔森相关系数度量特征之间的相关性, 则特征 $i$ 的独立性 $ind_i$ 定义为公式(18):

$$ind_i=\sum_{k=1,\ldots,m,\,k\neq i}(1-|r_{i,k}|) \tag{18}$$

第 $i,k(i,k=1,\ldots,m)$特征之间的皮尔森相关系数 $r_{i,k}$ 为公式(19), 其中, $f_{ji}$ 为样本 $j$ 在特征 $i$ 的取值, $\overline{f_i}$ 为特征 $i$ 在全部样本的均值:

$$r_{i,k}=\frac{\sum_{j=1}^{n}(f_{ji}-\bar{f}_i)(f_{jk}-\bar{f}_k)}{\sqrt{\sum_{j=1}^{n}(f_{ji}-\bar{f}_i)^2(f_{jk}-\bar{f}_k)^2}} \tag{19}$$

该定义保障了与其余特征相关性越小的特征, 其独立性越强.

**定义 10(特征重要度)**. 为了量化特征对于分类的贡献,定义第 $i$ 个特征的重要度为式(20)的 $score_i$, 即特征 $i$ 的辨识度 $dis_i$ 与其独立性 $ind_i$ 的乘积, $score_i$ 越大, 第 $i$ 个特征越重要.

$$score_i=dis_i\times ind_i \tag{20}$$

该定义不仅可以保障辨识度和独立性均大的特征对分类的贡献大, 即重要度高, 同时还可以在辨识度与独立性之间进行一定的折中, 以便对分类贡献比较大的特征都能被选择到.

### 3.2 算法描述

设数据集 $\boldsymbol{Data}\in\Re^{n\times(m+1)}$, 即 ***Data*** 包含 $n$ 个样本, 每个样本有 $m$ 个特征, 第 $i(i=1,\ldots,m)$个特征向量 $\boldsymbol{f}_i\in\Re^n$, $\boldsymbol{Y}\in\Re^{n\times 1}$ 为类标向量. 则 DFSIP 算法的详细步骤描述如下.

输入: 训练集 ***X***;

输出: 特征子集 ***F***;

**Begin**

根据定义 5 构建可辨识矩阵 $\boldsymbol{M}\in\Re^{n\times n}$;

置候选特征子集 ***CS*** 为原始特征子集 ***C***, 特征子集 ***FS*** 为空集，即置 ***FS***=∅;

**For** $i$=1 to ||***CS***|| **do**

根据公式(17)计算 $dis_i$;

根据公式(18)计算 $ind_i$;

根据公式(20)计算 $score_i$;

**End** //**of For**

**While** ***CS***≠∅ **do**

将当前 ***CS*** 的所有特征散列在以辨识度为横坐标、独立性为纵坐标的 2D 空间;

$score_{\max}$=min;

**For** $i$=1 to ||***CS***|| **do**

**If** $score_i \succ score_{\max}$ **then**

$score_{\max}=score_i$;

$f_{\max}=f_i$;

**End** // **of if**

**End** // **of For**

***FS***=***FS***∪{$f_{\max}$};

根据定义 6, 使用 $f_{\max}$ 对可辨识矩阵 ***M*** 进行约简, 更新 ***M*** 为 ***M′***;

根据定义 7, 使用新可辨识矩阵 ***M′*** 更新候选特征子集 ***CS***;

**End** // **of While**;

输出得到的特征子集 ***FS***;

**End**

基于 ***FS*** 中特征构造 *K*-ELM 分类器, 以其分类性能评价 ***FS*** 的类别辨别能力. 采用网格搜索选择 *K*-ELM 的惩罚参数 $C$ 和核函数参数 $\gamma$ 的最佳值. 实验中, 参数搜索空间分别为 $\log_2 C\in\{-18,-17,\ldots,15\}$ 和 $\log_2\gamma\in\{-18,-17,\ldots,15\}$, 搜索步长为 2 倍速. 以 *K*-ELM 分类性能评价特征子集 ***FS*** 的分类能力, 进而评价特征选择算法 DFSIP 的性能.

### 3.3 算法复杂度分析

本文 DFSIP 算法的空间复杂度主要是可辨识矩阵占用的空间, 可辨识矩阵 $\boldsymbol{M}$ 是一个 $n\times n$ 矩阵, 其元素 $m_{ij}$ 是对象 $x_i$ 与 $x_j$ 在取值上距离超过阈值$\delta$的属性构成的集合, 最多不超过总属性个数 $m$. 因此, 空间复杂度不超过 $O(mn^2)$. 对于高维小样本的基因表达数据集, 其特征数远大于样本数, 即 $m>>n$. DFSIP 算法的时间复杂度主要耗费在建立初始可辨识矩阵, 计算属性的辨识度、独立性和重要度, 选择当前最重要的属性, 约简更新矩阵, 重构候选特征子集. 其中, 建立初始可辨识矩阵的时间复杂度不超过 $O(mn^2)$; 计算每个属性的辨识度、独立性和重要度(对分类的贡献)的时间复杂度均不超过 $O(mn)$, 因此, 计算所有特征的辨识度、独立性和重要度的时间复杂度不超过 $O(m^2n)$; 选择最重要的属性的时间复杂度不超过为 $O(m)$, 约简更新可辨识矩阵的时间复杂度不超过 $O(n^2)$, 重构新候选特征子集的时间复杂度不超过 $O(n^2)$. 因此, 本文 DFSIP 算法的总时间复杂度上限为 $O(mn^2+m^2n+||\boldsymbol{CS}||\times(m+n^2+n^2))\prec O(mn^2+m^2n+m\times(m+n^2+n^2))=O(3mn^2+(n+1)m^2)\approx O(m^2n)$.

本文实验的对比算法 FSIP 的时间复杂度约为 $O(m^2n)$, mRMR 的时间复杂度约为 $O(m^2)$, LLE Score 的时间复杂度约为 $O(n^2)$, DRJMIM 的时间复杂度约为 $O(mn)$, AVC 算法的时间复杂度约为 $O(m^2)$, AMID 算法的时间复杂度约为 $O(mn)$. 与各种对比算法相比, 本文 DFSIP 算法不具有时间上的优势.

### 3.4 算法扩展性分析

本文完全自适应的特征选择算法 DFSIP 不仅适用于静态数据, 还适用于流式数据. 假设数据集 $\boldsymbol{X}\in\mathfrak{R}^{n\times m}$, 则可辨识矩阵 $\boldsymbol{M}\in\mathfrak{R}^{n\times n}$, 新来 $l$ 个样本, 则数据集为 $\boldsymbol{X}_{new}\in\mathfrak{R}^{(n+l)\times m}$, 可辨识矩阵变成 $\boldsymbol{M}_{new}\in\mathfrak{R}^{(n+l)\times(n+l)}$. 假设数据集 $\boldsymbol{X}$, $\boldsymbol{X}_{new}$ 的特征子集分别为 $\boldsymbol{FS}$ 和 $\boldsymbol{FS}_{new}$, 则 $\boldsymbol{M}$, $\boldsymbol{M}_{new}$ 的关系是 $\boldsymbol{M}_{new}=\begin{bmatrix}\boldsymbol{M}_{n\times n} & \boldsymbol{B}_{n\times l}\\ \boldsymbol{A}_{l\times n} & \boldsymbol{D}_{l\times l}\end{bmatrix}$. 使用由可辨识矩阵 $\boldsymbol{M}$ 得到的特征子集 $\boldsymbol{S}$ 中包含的特征对可辨识矩阵 $\boldsymbol{M}_{new}$ 的块矩阵 $\boldsymbol{B}_{n\times l}$, $\boldsymbol{A}_{l\times n}$, $\boldsymbol{D}_{l\times l}$ 进行约简, 若该 3 个块矩阵均被约简为空矩阵, 则新数据集 $\boldsymbol{X}_{new}\in\mathfrak{R}^{(n+l)\times m}$ 的特征子集 $\boldsymbol{S}_{new}$ 与原数据集的特征子集 $\boldsymbol{S}$ 完全相同, $\boldsymbol{S}_{new}=\boldsymbol{S}$; 否则, 新数据集 $\boldsymbol{X}_{new}\in\mathfrak{R}^{(n+l)\times m}$ 的特征子集 $\boldsymbol{S}_{new}$ 中的元素可能增加, 即$||\boldsymbol{S}_{new}||\succ||\boldsymbol{S}||$, 但求解新数据集 $\boldsymbol{X}_{new}\in\mathfrak{R}^{(n+l)\times m}$ 的特征子集不需要重新开始, 可以在特征子集 $\boldsymbol{S}$ 的基础上求解新数据集的特征子集 $\boldsymbol{S}_{new}$. 特征子集 $\boldsymbol{S}$ 和 $\boldsymbol{S}_{new}$ 之间存在如下关系: $\boldsymbol{S}\subseteq\boldsymbol{S}_{new}$.

## 4 实验结果与分析

本文实验分为两部分: 第 1 部分为参数敏感度分析, 探索阈值参数$\delta$与选择的特征数量、选择的特征子集的分类能力等的相关性, 并为第 2 部分实验选择最佳阈值参数$\delta$; 第 2 部分将比较本文算法 DFSIP 与特征选择算法 FSIP[37], mRMR[46], LLE Score[47], DRJMIM[48], AVC[49], AMID[50]的性能.

实验为 5-折交叉验证实验, 为获得随机实验数据, 实验前对数据进行打乱, 并使用最大最小标准化预处理实验数据. 编程语言为 MatlabR2017b, 实验环境为 Win10, 64bit OS, 32GB RAM, Intel(R) Xeon(R) E-2186M CPU@2.90GHz 2.90GHz.

### 4.1 评价指标

使用 5-折交叉验证的平均预测准确率 Accuracy、查全率 Recall、查准率 Precision、$F$-measure、$F2$-measure[50]、AUC[2,51,52]和 MCC[53]评价各特征选择算法发现的特征子集的分类辨识能力.

### 4.2 实验数据

使用 7 个常用的基因数据集 Colon, CNS, GLIOMA, Carcinom, Gas, leukemia, prostate 来测试 DFSIP 算法. 数据集详细细信息见表 1. 该 7 个实验数据集获得地址参见文献[37].

表 1 实验用基因数据集描述

| Datasets | #Samples | #Features | #Classes |
|---|---|---|---|
| Colon | 62 | 2 000 | 2 |
| CNS | 90 | 7 129 | 2 |
| GLIOMA | 50 | 4 434 | 4 |
| Carcinom | 174 | 9 182 | 11 |
| Gas | 65 | 22 645 | 2 |
| leukemia | 72 | 7 129 | 2 |
| prostate | 136 | 12 601 | 2 |

### 4.3 阈值参数与特征子集分类能力及规模的相关性

为探索阈值参数$\delta$对本文 DFSIP 算法性能的影响, 本节以表 1 的 Colon 数据集为例, 采用 5-折交叉验证实验, 通过改变阈值$\delta$重复 5-折交叉验证, 并再重新划分数据集进行 5-折交叉验证实验. 重复 5 次, 求均值, 比较 DFSIP 算法选特的特征子集的分类准确率、特征子集规模(即特征子集包含的特征数), 研究阈值$\delta$与 DFSIP 算法选择的特征子集的分类能力、包含特征数量的关系. 分类器使用核函数为 RBF[54]的 *K*-ELM. 图 1 展示了阈值$\delta$与特征数和特征子集分类准确率的关系.

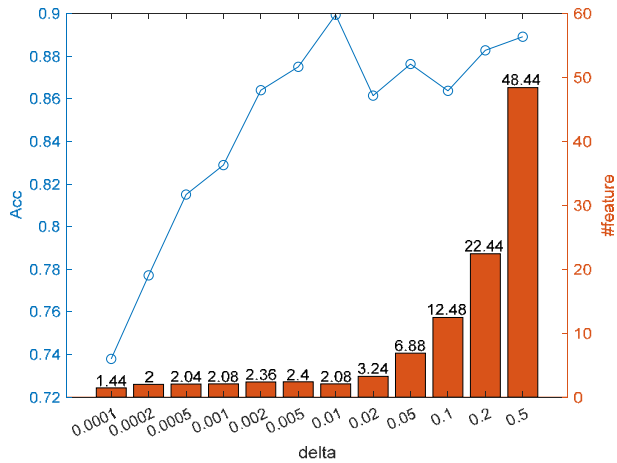
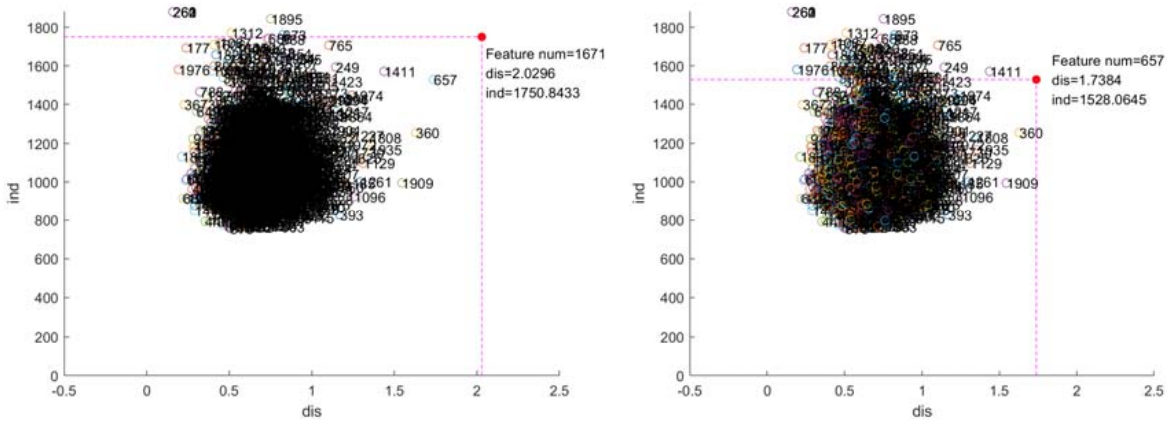



图 1 阈值$\delta$与特征数量和特征子集分类准确率的关系

从图 1 结果可以看出: 本文 DFSIP 算法选择的特征子集的分类能力开始时是随着邻域粗糙集阈值$\delta$的增大而上升, 即 DFSIP 算法选择的特征子集的分类能力与阈值$\delta$在开始时正相关; 但当阈值$\delta$上升到一定程度, 再加大阈值$\delta$反而会使 DFSIP 算法选择的特征子集的分类能力下降, 并出现波动. 分析原因是: 当阈值$\delta$较小时, DFSIP 算法容易陷入过度约简, 导致选择不到有用的特征; 而当阈值过大时, 则会导致不能排除无用的特征. 图 1 显示: colon 数据集在阈值$\delta$取 0.01 时, 选择到的特征子集不仅得到了最好的分类效果, 也包含较少数目特征. 后面实验对各数据集通过实验选择合适的阈值参数$\delta$.

### 4.4 DFSIP算法性能测试

本节将对比本文 DFSIP 算法与已有 FSIP 算法, 以及经典的 mRMR, LLE Score, DRJMIM, AVC, AMID 算法的性能. 图 2 展示了在 Colon 数据集某一折上的特征选择过程. 表 2 展示了各对比算法的 5 折交叉验证的平均预测准确率 Accuracy、查准率 Precision、查全率 Recall、*F*-measure、*F*2-measure、AUC 以及 MCC, 其中, 加粗表示最优结果. 对比算法中的参数均采用默认值, 并保持最后选择的特征数量与 DFIP 选择的一致.




(a) 第 1 个最重要特征 (b) 第 2 个最重要特征

图 2 Colon 数据集 5 折划分的某一折的特征选择过程

表 2 各算法 5 折交叉验证的平均实验结果

| Datasets | Algorithms | Accuracy | Recall | AUC | Precision | *F*-measure | *F*2-measure | MCC | #Features | $\delta$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Colon | DFSIP | **0.901 3** | **0.975 0** | **0.903 8** | **0.911 1** | **0.934 9** | 0.766 0 | **0.735 4** | 2 | 0.01 |
| | FSIP | 0.855 1 | 0.925 0 | 0.833 8 | 0.872 8 | 0.890 2 | **0.871 0** | 0.727 0 | | |
| | mRMR | 0.855 1 | 0.950 0 | 0.793 8 | 0.847 8 | 0.893 8 | 0.866 9 | 0.733 0 | | |
| | LLE Score | 0.837 2 | **0.975 0** | 0.805 0 | 0.823 9 | 0.889 2 | 0.715 8 | 0.632 0 | | |
| | DRJMIM | 0.756 4 | 0.9500 | 0.697 5 | 0.754 8 | 0.836 5 | 0.647 6 | 0.448 1 | | |
| | AVC | 0.771 8 | **0.975 0** | 0.787 5 | 0.757 6 | 0.849 8 | 0.677 8 | 0.485 1 | | |
| | AMID | 0.710 3 | **0.975 0** | 0.610 0 | 0.699 4 | 0.813 6 | 0.458 2 | 0.291 6 | | |
| CNS | DFSIP | **0.900 0** | 0.966 7 | 0.844 4 | **0.894 1** | **0.927 9** | **0.910 4** | **0.809 2** | 3.8 | 0.005 |
| | FSIP | 0.855 6 | 0.966 7 | 0.855 6 | 0.843 1 | 0.899 6 | 0.876 5 | 0.721 0 | | |
| | mRMR | 0.877 8 | 0.966 7 | 0.866 7 | 0.866 8 | 0.913 2 | 0.892 9 | 0.767 7 | | |
| | LLE Score | 0.811 1 | **1.000 0** | 0.813 9 | 0.782 6 | 0.877 1 | 0.877 1 | 0.643 3 | | |
| | DRJMIM | 0.822 2 | 0.983 3 | 0.788 9 | 0.807 1 | 0.883 9 | 0.860 6 | 0.633 0 | | |
| | AVC | 0.888 9 | 0.983 3 | **0.894 4** | 0.878 1 | 0.925 4 | 0.902 1 | 0.775 4 | | |
| | AMID | 0.755 6 | 0.933 3 | 0.675 0 | 0.771 7 | 0.836 2 | 0.644 3 | 0.426 2 | | |
| GLIOMA | DFSIP | 0.621 4 | 0.672 2 | 0.860 2 | 0.665 0 | 0.650 2 | 0.504 0 | 0.457 9 | 3.4 | 0.000 1 |
| | FSIP | 0.615 0 | 0.455 6 | **0.920 4** | 0.502 8 | 0.461 9 | 0.335 7 | 0.274 9 | | |
| | mRMR | **0.729 5** | **0.694 4** | 0.857 9 | **0.783 3** | **0.719 4** | **0.659 5** | **0.575 3** | | |
| | LLE Score | 0.416 8 | 0.438 9 | 0.819 4 | 0.433 3 | 0.417 9 | 0.143 3 | 0.088 9 | | |
| | DRJMIM | 0.697 3 | 0.655 6 | 0.877 8 | 0.622 8 | 0.629 1 | 0.494 6 | 0.451 9 | | |
| | AVC | 0.501 8 | 0.455 6 | 0.819 4 | 0.533 3 | 0.468 3 | 0.369 7 | 0.283 0 | | |
| | AMID | 0.448 6 | 0.505 6 | 0.850 0 | 0.470 0 | 0.459 9 | 0.167 1 | 0.114 9 | | |
| Carcinom | DFSIP | **0.641 2** | **0.685 6** | 0.879 4 | **0.665 2** | **0.672 3** | 0.316 6 | 0.310 0 | 4 | 0.01 |
| | FSIP | 0.574 2 | 0.545 0 | 0.902 2 | 0.523 5 | 0.530 7 | 0.273 9 | 0.262 0 | | |
| | mRMR | 0.570 7 | 0.627 0 | 0.862 9 | 0.616 6 | 0.616 5 | **0.349 7** | **0.330 1** | | |
| | LLE Score | 0.359 4 | 0.434 1 | 0.826 4 | 0.397 5 | 0.410 3 | 0.090 7 | 0.076 3 | | |
| | DRJMIM | 0.342 1 | 0.525 6 | 0.789 7 | 0.470 3 | 0.490 4 | 0.123 8 | 0.111 9 | | |
| | AVC | 0.329 8 | 0.283 0 | 0.899 8 | 0.274 6 | 0.275 6 | 0.058 0 | 0.049 1 | | |
| | AMID | 0.188 0 | 0.118 5 | **0.902 6** | 0.087 1 | 0.099 6 | 0.015 8 | 0.012 3 | | |
| Gas | DFSIP | 0.938 3 | **0.900 0** | 0.916 0 | 0.966 7 | **0.927 3** | 0.943 4 | **0.854 1** | 3.4 | 0.01 |
| | FSIP | **0.939 6** | 0.866 7 | 0.934 5 | **1.000 0** | **0.927 3** | **0.948 2** | 0.824 8 | | |
| | mRMR | 0.888 5 | 0.780 0 | 0.8219 | 0.960 0 | 0.841 6 | 0.906 2 | 0.720 8 | | |
| | LLE Score | 0.848 4 | 0.766 7 | 0.816 7 | 0.900 0 | 0.800 0 | 0.864 2 | 0.671 0 | | |
| | DRJMIM | 0.818 7 | 0.833 3 | 0.779 2 | 0.791 0 | 0.806 6 | 0.821 3 | 0.660 7 | | |
| | AVC | 0.890 8 | 0.786 7 | **0.938 6** | 0.971 4 | 0.849 8 | 0.912 6 | 0.731 8 | | |
| | AMID | 0.720 3 | 0.593 3 | 0.727 9 | 0.776 2 | 0.618 3 | 0.742 9 | 0.420 4 | | |

表 2 各算法 5 折交叉验证的平均实验结果(续)

| Datasets | Algorithms | Accuracy | Recall | AUC | Precision | *F*-measure | *F*2-measure | MCC | #Features | $\delta$ |
|---|---|---|---|---|---|---|---|---|---|---|
| leukemia | DFSIP | **0.958 1** | **1.000 0** | 0.975 1 | 0.941 8 | **0.969 4** | **0.969 4** | **0.936 7** | 4.4 | 0.01 |
| | FSIP | **0.958 1** | **1.000 0** | 0.953 8 | 0.941 8 | **0.969 4** | **0.969 4** | **0.936 7** | | |
| | mRMR | 0.918 1 | **1.000 0** | 0.920 4 | 0.895 7 | 0.943 3 | 0.943 3 | 0.863 1 | | |
| | LLE Score | 0.930 5 | 0.957 8 | **0.978 7** | 0.937 8 | 0.947 3 | 0.927 3 | 0.856 7 | | |
| | DRJMIM | 0.764 8 | 0.980 0 | 0.734 2 | 0.744 4 | 0.845 5 | 0.805 1 | 0.529 0 | | |
| | AVC | 0.945 7 | 0.957 8 | 0.962 2 | **0.966 7** | 0.959 5 | 0.945 5 | 0.881 5 | | |
| | AMID | 0.792 4 | 0.955 6 | 0.769 3 | 0.796 7 | 0.858 4 | 0.850 1 | 0.600 4 | | |
| prostate | DFSIP | **0.933 6** | 0.920 8 | **0.956 9** | **0.962 4** | **0.936 9** | **0.935 9** | **0.859 3** | 7 | 0.01 |
| | FSIP | 0.816 6 | 0.880 8 | 0.845 7 | 0.829 8 | 0.848 1 | 0.830 3 | 0.651 5 | | |
| | mRMR | 0.860 2 | 0.921 7 | 0.856 5 | 0.856 9 | 0.883 2 | 0.873 8 | 0.753 8 | | |
| | LLE Score | 0.888 8 | 0.935 0 | 0.925 1 | 0.883 1 | 0.905 9 | 0.895 1 | 0.800 1 | | |
| | DRJMIM | 0.721 6 | **0.936 7** | 0.730 0 | 0.688 6 | 0.791 7 | 0.763 4 | 0.532 8 | | |
| | AVC | 0.904 5 | 0.921 7 | 0.898 5 | 0.914 5 | 0.916 3 | 0.905 5 | 0.811 2 | | |
| | AMID | 0.684 5 | 0.947 5 | 0.679 0 | 0.672 5 | 0.776 8 | 0.620 1 | 0.408 0 | | |

图 2 显示: 在该折上对由所有特征构成的候选特征子集的特征, 采用公式(17)、公式(18)和公式(20)分别计算特征的辨识度、独立性和重要度, 将所有特征展示在分别以辨识度与独立性为横、纵坐标的 2D 空间(如图 2(a)所示), 选择最右上角的特征, 即第 1 个最重要特征, 编号 1671. 然后进入可辨识矩阵的约简过程, 根据定义 6 对可辨识矩阵进行约简, 使可辨识矩阵中与特征 1671 有交集的元素置空, 可辨识矩阵中的非空元素的并集构成新的候选特征子集, 将候选特征子集中余下特征展示在以辨识度为横坐标、独立性为纵坐标的 2D 空间(如图 2(b)所示), 选出其中最重要的特征, 即最右上角的编号为 657 的特征. 然后再进入可辨识矩阵的约简过程, 依据定义 6 对可辨识矩阵再次进行约简, 使可辨识矩阵中与特征 657 有交集的元素置空, 即通过置空包含特征 657 的可辨识矩阵的元素, 剔除掉与特征 657 相关的冗余特征, 再将可辨识矩阵中的非空元素作并集, 构成新的候选特征子集, 将候选特征子集中的特征再展示在以辨识度、独立性分别作横、纵坐标的 2D 空间, 从中选择最重要的特征. 如下迭代, 直到候选特征子集为空集, 则完成了该折的特征选择过程. 实际上, 使用第 2 个最重要特征 657 约简可辨识矩阵后, 可辨识矩阵则变为空矩阵.

因此, 图 2 显示的 Colon 数据集的 5-折交叉验证实验的某折数据的特征选择过程, 在选择到第 2 个最重要特征 657 后, 使用该特征对可辨识矩阵进行约简, 约简的结果使得可辨识矩阵的每个元素成为了空集, 从而使得候选特征子集变成了空集, DFSIP 算法收敛, 得到该折划分对应的特征子集 $\boldsymbol{FS}=\{f_{1671},f_{657}\}$.

表 2 的实验结果显示.

- 在 Colon 数据集, 本文 DFSIP 算法在 Accuracy, AUC, Precision, *F*-measure, MCC 指标上要优于所有对比算法, 在 Recall 指标上与 LLE Score, AVC, AMID 算法持平且优于其他对比算法, 在 *F*2-measure 指标上劣于 FSIP, mRMR 算法但优于其他对比算法;
- 在 CNS 数据集, 本文 DFSIP 算法的 Accuracy, Precision, *F*-measure, *F*2-measure 和 MCC 指标均优于对比算法; Recall 指标与 FSIP, mRMR 持平, 仅仅优于 AMID 算法, 劣于对比算法 LLE Score, DRJMIM 和 AVC; LLE Score 算法在 CNS 数据集的 Recall 达到最优值 1; AVC 算法在 CNS 数据集的 AUC 指标最佳, 为 0.8944;
- 在 GLIOMA 数据集, 本文 DFSIP 算法在 Accuracy 指标上劣于 mRMR, DRJMIM 算法但优于其他对比算法, 在 AUC 指标上劣于 FSIP 和 DDRJMIM 算法但优于其他对比算法, 在 Recall, Precision, *F*-measure, *F*2-measure, MCC 指标上劣于 mRMR 算法但优于其他对比算法. 总体来看, 在 GLIOMA 数据集, mRMR 算法的性能最优, 其次是本文提出的 DFSIP 算法;
- 在 Carcinom 数据集, 本文 DFSIP 算法在 Accuracy, Recall, Precision, *F*-measure 指标上都优于对比算法, 在 AUC 指标上劣于 FSIP, AVC, AMID 算法但优于其他对比算法, 在 *F*2-measure, MCC 指标上劣于 mRMR 算法但优于其他对比算法;
- 在 Gas 数据集, 本文 DFSIP 算法在 Accuracy 指标上劣于 FSIP 算法但优于其他对比算法; 在 Recall, MCC 指标上优于所有对比算法; 在 AUC 指标上劣于 FSIP, AVC 算法但优于其他对比算法; 在

Precision 指标上劣于 FSIP, AVC 算法但优于其他对比算法, FSIP 此时的 precision 为最优值 1; 在 *F*-measure 指标上与 FSIP 算法持平且优于其他对比算法; 在 *F*2-measure 指标上劣于 FSIP 算法但优于其他对比算法;

- 在 leukemia 数据集, 本文 DFSIP 算法在 Accuracy, *F*-measure, *F*2-measure, MCC 指标上都与 FSIP 算法持平且优于其他对比算法, 在 Recall 指标上与 FSIP 和 mRMR 持平且优于其他对比算法, 在 AUC 指标上仅劣于 LLE Score 算法但优于其他对比算法, 在 Precision 指标上劣于 AVC 算法但与 FSIP 算法持平且优于其他对比算法;
- 在 prostate 数据集, 本文 DFSIP 算法在 Accuracy, AUC, Precision, *F*-measure, *F*2-measure, MCC 指标上均优于所有对比算法; 在 Recall 指标上仅优于 FSIP 算法, 劣于其他比较算法 mRMR, LLE Score, DRJMIM, AVC 和 AMDI.

综上分析可见: 本文提出的完全自适应的特征选择算法 DFSIP 在针对高维基因数据集进行特征选择时, 可以选择出分类识别能力很好的特征子集, 实现数据降维目的.

### 4.5 统计重要性检测

本节检验本文 DFSIP 算法与对比算法 FSIP, mRMR, LLE Score, DRJMIM, AVC, AMID 是否具有统计意义上的显著性差异. 首先采用 Friedman 检验来检测该 7 个算法间是否存在统计意义上的显著性差异[2,55]. 在 Friedman 检验检测到算法间存在显著性差异后, 采用多重比较考察算法两两之间是否有显著性不同. 表 3 展示了各算法特征子集的 *K*-ELM 分类器的评价指标 Accuracy, Recall, AUC, precision, *F*-measure, *F*2-measure 以及 *MCC* 在 $\alpha$=0.05 时 Friedman 检测的结果.

表 3 各算法所选特征子集分类能力的 Friedman 检测结果

| | Accuracy | Recall | AUC | Precision | *F*-measure | *F*2-measure | MCC |
|---|---|---|---|---|---|---|---|
| $\chi^2$ | 26.815 4 | 5.723 6 | 15.329 9 | 26.163 7 | 28.246 2 | 26.347 8 | 29.140 7 |
| *df* | 6.000 0 | 6.000 0 | 6.000 0 | 6.000 0 | 6.000 0 | 6.000 0 | 6.000 0 |
| *p* | 0.000 2 | 0.454 9 | 0.017 8 | 0.000 2 | 0.000 1 | 0.000 2 | 0.000 1 |

表 3 各算法的 Friedman 检测结果显示, $p$<0.05 对 Recall 指标不成立, 但对 Accuracy, AUC, Precision, *F*-measure, *F*2-measure 和 MCC 指标均成立. 说明在 $p$<0.05 时, 各算法选择的特征子集在 Recall 指标上无显著性区别. 但从各算法选择的特征子集对应分类器的 Accuracy, AUC, Precision, *F*-measure, *F*2-measure 和 MCC 指标来看, 各算法存在统计意义上的显著不同. 这也说明这些算法选择的特征子集对正类的识别能力相仿, 对负类的识别能力差异很大.

因此, 需要对各算法依据其特征子集的 *K*-ELM 分类器的 Accuracy, AUC, Precision, *F*-measure, *F*2-measure 以及 MCC 进一步采用多重比较检测各特征选择算法两两之间是否有统计意义上的显著性区别. 表 4–表 9 给出了在 0.95 可信水平下, 每两个算法在不同指标下的多重比较检验的结果. 表中上三角表示对应两算法的平均等级差, 下三角表示两算法的统计重要性, *表示相应的两算法在统计意义上存在显著性区别.

表 4 7 种特征选择算法结合 *K*-ELM 模型的 Accuracy 等级比较

| Algorithms | DFSIP | FSIP | mRMR | LLE Score | DRJMIM | AVC | AMID |
|---|---|---|---|---|---|---|---|
| DFSIP | | 1.357 1 | 1.714 3 | 3.214 3 | 3.785 7 | 2.214 3 | 5.214 3 |
| FSIP | | | 0.357 1 | 1.857 1 | 2.428 6 | 0.857 1 | 3.857 1 |
| mRMR | | | | 1.5 | 2.071 4 | 0.5 | 3.5 |
| LLE Score | | | | | 0.571 4 | −1 | 2 |
| DRJMIM | * | | | | | −1.571 4 | 1.428 6 |
| AVC | | | | | | | 3 |
| AMID | * | * | * | | | | |

表 5 7 种特征选择算法结合 *K*-ELM 模型的 AUC 等级比较

| Algorithms | DFSIP | FSIP | mRMR | LLE Score | DRJMIM | AVC | AMID |
|---|---|---|---|---|---|---|---|
| DFSIP | | 0.142 9 | 1.428 6 | 1.5 | 3.142 9 | 0.642 9 | 3.142 9 |
| FSIP | | | 1.285 7 | 1.357 1 | 3 | 0.5 | 3 |
| mRMR | | | | 0.071 4 | 1.714 3 | −0.785 7 | 1.714 3 |
| LLE Score | | | | | 1.642 9 | −0.857 1 | 1.642 9 |
| DRJMIM | | | | | | −2.5 | 0 |
| AVC | | | | | | | 2.5 |
| AMID | | | | | | | |

表 6 7 种特征选择算法结合 *K*-ELM 模型的 Precision 等级比较

| Algorithms | DFSIP | FSIP | mRMR | LLE Score | DRJMIM | AVC | AMID |
|---|---|---|---|---|---|---|---|
| DFSIP | | 1.571 4 | 1.5 | 3.214 3 | 3.642 9 | 1.5 | 5.071 4 |
| FSIP | | | −0.071 4 | 1.642 9 | 2.071 4 | −0.071 4 | 3.5 |
| mRMR | | | | 1.714 3 | 2.142 9 | 0 | 3.571 4 |
| LLE Score | | | | | 0.428 6 | −1.714 3 | 1.857 1 |
| DRJMIM | * | | | | | −2.142 9 | 1.428 6 |
| AVC | | | | | | | 3.571 4 |
| AMID | * | * | * | | | * | |

表 7 7 种特征选择算法结合 *K*-ELM 模型的 *F*-measure 等级比较

| Algorithms | DFSIP | FSIP | mRMR | LLE Score | DRJMIM | AVC | AMID |
|---|---|---|---|---|---|---|---|
| DFSIP | | 1.714 3 | 2.714 3 | 2.857 1 | 4.857 1 | 2.714 3 | 3.142 9 |
| FSIP | | | 1 | 1.142 9 | 3.142 9 | 1 | 1.428 6 |
| mRMR | | | | 0.142 9 | 2.142 9 | 0 | 0.428 6 |
| LLE Score | * | | | | 2 | −0.142 9 | 0.285 7 |
| DRJMIM | * | | | | | −2.142 9 | −1.714 3 |
| AVC | | | | | | | 0.428 6 |
| AMID | * | * | * | | | | |

表 8 7 种特征选择算法结合 *K*-ELM 模型的 *F*2-measure 等级比较

| Algorithms | DFSIP | FSIP | mRMR | LLE Score | DRJMIM | AVC | AMID |
|---|---|---|---|---|---|---|---|
| DFSIP | | 1.285 7 | 0.928 6 | 2.928 6 | 3.642 9 | 1.785 7 | 4.928 6 |
| FSIP | | | −0.357 1 | 1.642 9 | 2.357 1 | 0.5 | 3.642 9 |
| mRMR | | | | 2 | 2.714 3 | 0.857 1 | 4 |
| LLE Score | | | | | 0.714 3 | −1.142 9 | 2 |
| DRJMIM | * | | | | | −1.857 1 | 1.285 7 |
| AVC | | | | | | | 3.142 9 |
| AMID | * | * | * | | | | |

表 9 7 种特征选择算法结合 *K*-ELM 模型的 MCC 等级比较

| Algorithms | DFSIP | FSIP | mRMR | LLE Score | DRJMIM | AVC | AMID |
|---|---|---|---|---|---|---|---|
| DFSIP | | 2 | 1.357 1 | 3.5 | 4.071 4 | 2.214 3 | 5.357 1 |
| FSIP | | | −0.642 9 | 1.5 | 2.071 4 | 0.214 3 | 3.357 1 |
| mRMR | | | | 2.142 9 | 2.714 3 | 0.857 1 | 4 |
| LLE Score | * | | | | 0.571 4 | −1.285 7 | 1.857 1 |
| DRJMIM | * | | | | | −1.857 1 | 1.285 7 |
| AVC | | | | | | | 3.142 9 |
| AMID | * | | * | | | | |

表 4 基于各算法选择的特征子集的 *K*-ELM 分类器的 Accuracy 多重比较表明, 本文 DFSIP 算法与 DRJMIM 和 AMID 算法存在显著性不同. 我们之前研究提出 AMID 算法与 DFSIP, FSIP 和 mRMR 均存在显著性差异. DFSIP 与其他算法的差异不是统计意义上的显著性差异, 但是都存在差异.

表 5 基于各算法 AUC 的多重比较显示, 本文提出的 DFSIP 算法与其他对比算法间不存在统计意义上的显著性不同. 但可看出: DFSIP 与 FSIP, AVC 算法的等级差大于 0, 与算法 mRMR, LLE Score 的等级差大于 1, 与算法 DRJMIM, AMID 等级差大于 3. 该结果说明本文提出的 DFSIP 算法与对比算法 FSIP, AVC, MRMR, LLE Score, DRJMIM 和 AMID 存在不同程度的差异, 与 DRJMIM 和 AMID 算法的差异最大, 尽管这些差异还不构成显著性差异.

表 6 基于各算法选择的特征子集对应 *K*-ELM 分类器的 Precision 多重比较显示, 本文提出 DFSIP 算法与对比算法 DRJMIM, AMID 间存在显著性不同. 我们前期研究中提出的特征选择算法 AMID 与本文提出的 DFSIP 算法, 以及现有的 FSIP, mRMR 和 AVC 显著不同.

表 7 基于各算法选择的特征子集的 *K*-ELM 分类器的 *F*-measure 的多重比较显示, 本文提出的 DFSIP 算法与对比算法 LLE Score, DRJMIM, AMID 算法存在显著性不同. 我们前期研究的特征选择算法 AMID 与 DFSIP, FSIP 和 mRMR 存在显著性不同.

表 8 基于各算法对应特征子集的 *K*-ELM 分类器的 *F*2-measure 的多重比较表明, 本文 DFSIP 算法与对比的 DRJMIM 和 AMID 算法存在显著性不同. AMID 与 DFSIP, FSIP 和 mRMR 存在显著性不同. DFSIP 算法与其他算法间的差异尽管不是显著性差异, 但也均存在差异.

表 9 基于各算法对应特征子集的 K-ELM 分类器的 MCC 多重比较显示, 提出的 DFSIP 算法与对比算法 LLE Score, DRJMIM 和 AMID 存在显著不同. AMID 与 DFSIP 和 mRMR 存在显著性差异.

以上统计性检测分析表明. 本文提出的 DFSIP 算法与对比算法 LLE Score, DRJMIM, AMID 间存在统计意义上的显著性区别. 与 FSIP 算法间的差异在统计意义上不是显著的. 但基于各指标的多重检验说明, DFSIP 与 FSIP 存在差异, 只是差异不是显著的. 这验证了本文 DFSIP 算法不仅能实现特征子集的完全自适应选择, 而且选择的特征与现有特征选择算法 LLE Score, DRJMIM, AMID 选择的特征子集具有统计意义上的显著性区别, 且选择的特征子集的分类能力优于 FSIP 算法及对比算法选择的特征子集的分类能力.

### 4.6 各算法运行时间的比较

本节进一步比较本文提出的 DFSIP 算法与 6 个对比算法 FSIP, mRMR, LLE Score, DRJMIM, AVC 和 AMID 在各数据集的运行时间比较, 同时验证第 4.3 节关于算法时间复杂度的分析. 表 10 给出了各算法在表 1 所示的 7 个数据集 Colon, CNS, GLIOMA, Carcinom, Gas, leukemia 和 prostate 的 5 折交叉验证的平均运行时间比较.

表 10 各算法在表 1 数据集的 5 折交叉验证的平均运行时间比较 (s)

| Datasets | DFSIP | FSIP | mRMR | LLE Score | DRJMIM | AVC | AMID |
|---|---|---|---|---|---|---|---|
| Colon | 94.23±5.81 | 14.85±3.15 | 8.67±0.46 | 0.08±0.02 | 9.75±2.12 | 14.48±1.07 | 0.35±0.06 |
| CNS | 4349.30±348.20 | 141.73±15.24 | 10.26±0.19 | 0.46±0.04 | 43.81±9.75 | 16.22±3.07 | 2.26±0.44 |
| GLIOMA | 1198.70±352.96 | 62.67±5.66 | 9.94±0.18 | 0.23±0.18 | 26.73±7.06 | 41.17±5.18 | 1.29±0.37 |
| Carcinom | 6366.90±914.22 | 391.31±37.99 | 13.42±1.46 | 1.24±0.12 | 94.74±27.36 | 353.81±62.30 | 14.09±1.89 |
| Gas | 22528.00±5802.70 | 1516.50±243.05 | 12.21±1.43 | 2.98±0.04 | 83.00±7.80 | 7.72±0.50 | 7.07±2.07 |
| leukemia | 4035.90±1535.40 | 238.96±23.61 | 14.09±1.86 | 0.40±0.03 | 27.51±1.53 | 8.84±0.73 | 3.27±1.46 |
| prostate | 15462.00±3439.90 | 619.94±168.54 | 15.34±1.29 | 1.61±0.08 | 86.03±3.54 | 22.82±1.94 | 14.00±2.59 |

表 10 各算法的运行时间比较显示: 本文提出的 DFSIP 算法的时间效率最差, 第 2 费时的算法是 FSIP 算法, LLE Score 算法的时间效率最高, 接着是我们在之前研究中提出的 AMID 算法, mRMR, DRJMIM 和 AVC 算法的时间效率依次居中. 这些算法的实际运行时间与第 4.3 节的时间复杂度理论分析结果基本一致.

以上各算法的实际运行时间比较显示: 本文提出的 DFSIP 算法时间效率最差, 但其优点也很突出, 实现了特征子集的完全自动选择, 且选择的特征子集的分类识别能力最好, 这从表 2 的实验结果便可看出. 这体现了算法性能与运行效率之间的矛盾, 也就是用时间换取了性能提升. 探索时间效率高且能自动发现最优特征子集的特征选择算法, 是我们正在努力的方向, 也是需要该领域同行共同探索的问题.

## 5 结 论

提出了一种基于可辨识矩阵的完全自适应的 2D 特征选择新算法 DFSIP, 克服了 2D 特征选择算法 FSIP 需要人工参与发现最优特征子集的不足. DFSIP 算法以邻域信息增益定义特征辨识度, Pearson 相关系数定义特征独立性, 并以两者乘积作为特征重要程度的度量. 所有特征展示在分别以辨识度、独立性作横、纵坐标的 2D 空间, 每次从当前特征中选择最重要的特征加入特征子集, 并以此特征约简可辨识矩阵, 然后从可辨识矩阵的余下特征中选择最重要特征, 依次迭代, 直到余下的特征为空, 从而自适应地获得最优特征子集. 构造最优特征子集对应的 K-ELM 分类器, 以其性能评价最优特征子集的类别辨识能力, 进而衡量特征选择算法的

性能.

经典基因数据集的 5-折交叉验证实验, 以及与 FSIP 算法和经典特征选择算法 mRMR, LLE Score, DRJMIM, AVC, AMID 的实验比较和统计重要度检测分析发现: 提出的完全自适应的 2D 空间特征选择算法 DFSIP 是一种有效的特征选择算法, 可以选择出性能很好的特征子集. DFSIP 算法不仅完全实现了自适应的特征选择, 且使特征选择过程可视化, 并能用于任意维数据的特征选择和流式数据的特征选择, 实现数据的很好降维. DFSIP 的不足之处是: 构建邻域可辨识矩阵时的阈值需要实验确定, 且时间效率较差. 如何自动选择合适的邻域可辨识矩阵的阈值以及如何提高算法效率, 是我们未来的研究方向.

**References**:


[1] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997, 97(1-2): 245−271.

[2] Borg A, Lavesson N, Boeva V. Comparison of clustering approaches for gene expression data. In: Jaeger M, Nielsen TD, Viappiani P, eds. Proc. of the 12th Scandinavian Conf. on Artificial Intelligence. Amsterdam: IOS, 2013. 55−64.

[3] Chen HM, Li TR, Fan X, *et al*. Feature selection for imbalanced data based on neighborhood rough sets. Information Sciences, 2019, 483: 1−20.

[4] Gu SK, Cheng R, Jin YC. Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Computing, 2018, 22(3): 811−822.

[5] Gui J, Sun ZN, Ji SW, *et al*. Feature selection based on structured sparsity: A comprehensive study. IEEE Trans. on Neural Networks and Learning Systems, 2017, 28(7): 1490−1507.

[6] Xie JY, Lei JH, Xie WX, *et al*. Two-stage hybrid feature selection algorithms for diagnosing erythemato-squamous diseases. Health Information Science and Systems, 2013, 1: 10.

[7] Xie JY, Gao HC. Statistical correlation and *K*-means based distinguishable gene subset selection algorithms. Ruan Jian Xue Bao/ Journal of Software, 2014, 25(9): 2050−2075 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4644.htm [doi: 10.13328/j.cnki.jos.004644]

[8] Xie JY, Zhou Y. A new criterion for clustering algorithm. Journal of Shaanxi Normal University (Natural Science Edition), 2015, 43(6): 1−8 (in Chinese with English abstract). [doi: 10.15983/j.cnki.jsnu.2015.06.161]

[9] Yao X, Wang XD, Zhang YX, *et al*. Summary of feature selection algorithms. Control and Decision, 2012, 027(2): 161−166, 192 (in Chinese with English abstract).

[10] Almuallim H, Dietterich TG. Learning Boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 1994, 69(1-2): 279−305.

[11] Kira K, Rendell LA. A practical approach to feature selection. In: Sleeman D, Edwards P, eds. Proc. of the 9th Int'l Workshop (ML'92). California: Morgan Kaufmann Publishers, 1992. 249−256.

[12] Kononenko I. Estimating attributes: Analysis and extension of relief. In: Bergadano F, De Raedt L, eds. Proc. of the European Conf. on Machine Learning (ECML'94). Berlin: Springer, 1994. 171−182.

[13] Liu H, Motoda H. Feature Selection for Knowledge Discovery and Data Mining. New York: Kluwer Academic Publishers, 1998.

[14] Arauzo-Azofra A, Benitez JM, Castro JL. Consistency measures for feature selection. Journal of Intelligent Information Systems, 2008, 30(3): 273−292.

[15] Dash M, Liu H. Consistency-based search in feature selection. Artificial Intelligence, 2003, 151(1-2): 155−176.

[16] Almuallim H, Dietterich TG. Learning with many irrelevant features. In: Proc. of the 9th National Conf. on Artificial Intelligence, Vol.2. California: AAAI, 1991. 547−552.

[17] Liu H, Setiono R. A probabilistic approach to feature selection—A filter solution. In: Proc. of the 9th Int'l Conf. on Industrial and Engineering Applications of AI and ES, Vol.96. 1996. 319−327.

[18] Devijver PA, Kittler J. Pattern Recognition: A Statistical Approach. London: Prentice Hall Int'l, 1982.

[19] Hall MA. Correlation-Based feature subset selection for machine learning [Ph.D. Thesis]. Hamilton: The University of Waikato, 1998.

[20] Zhang DQ, Chen SC, Zhou ZH. Constraint score: A new filter method for feature selection with pairwise constraints. Pattern Recognition, 2008, 41(5): 1440−1451.

[21] Hall MA. Correlation-based feature selection of discrete and numeric class machine learning. In: Langley P, ed. Proc. of the 17th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000. 359−366.

[22] Ben-Bassat M. Pattern recognition and reduction of dimensionality. Handbook of Statistics, 1982, 2: 773−910.

[23] Finkelstein L. Optical pattern recognition. Optica Acta: Int'l Journal of Optics, 1980, 27(11): 1502−1502.

[24] Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. on Neural Networks, 1994, 5: 537−550.

[25] Kwak N, Choi CH. Input feature selection by mutual information based on Parzen window. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 24(12): 1667−1671.

[26] Huang JJ, Cai Y, Xu XM. A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognition Letters, 2007, 28(13): 1825−1844.

[27] Novovicová J, Somol P, Haindl M, *et al*. Conditional mutual information based feature selection for classification task. In: Rueda L, Mery D, Kittler J, eds. Proc. of the Progress in Pattern Recognition, Image Analysis and Applications (CIARP 2007). Berlin: Springer, 2007. 417−426.

[28] Qu GZ, Hariri S, Yousif M. A new dependency and correlation analysis for features. IEEE Trans. on Knowledge and Data Engineering, 2005, 17: 1199−1207.

[29] Estevez PA, Tesmer M, Perez CA, *et al*. Normalized mutual information feature selection. IEEE Trans. on Neural Networks, 2009, 20(2): 189−201.

[30] Liu HW. A study on feature selection algorithms using information entropy [Ph.D. Thesis]. Changchun: Jilin University, 2010 (in Chinese with English abstract).

[31] Fleuret F. Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research, 2004, 5(4941): 1531−1555.

[32] Reshef DN, Reshef YA, Finucane HK, *et al*. Detecting novel associations in large data sets. Science, 2011, 334(6062): 1518−1524.

[33] Li ZQ, Du JQ, Nie B, *et al*. Summary of feature selection methods. Computer Engineering and Applications, 2019, 55(24): 10−19 (in Chinese with English abstract). [doi: 10.3778/j.issn.1002-8331.1909-0066]

[34] He XF, Cai D, Niyogi P. Laplacian score for feature selection. In: Weiss Y, Schölkopf B, Platt JC, eds. Proc. of the 18th Int'l Conf. on Neural Information Processing Systems. Cambridge: MIT, 2005. 507−514.

[35] Bishop CM. Neural Networks for Pattern Recognition. Oxford: Oxford University Press, 1995.

[36] Van'T Veer LJ, Dai H, Van De Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415(6871): 530−536.

[37] Xie JY, Wu ZZ, Zheng QQ. An adaptive 2D feature selection algorithm based on information gain and Pearson correlation coefficient. Journal of Shaanxi Normal University (Natural Science Edition), 2020, 48(6): 69−81 (in Chinese with English abstract). [doi: 10. 15983/j.cnki.jsnu.2020.01.019]

[38] Mao F. College Physics II. Wuhan: Huazhong University of Science and Technology Press, 2010 (in Chinese).

[39] Pawlak Z. Rough sets. Int'l Journal of Computer & Information Sciences, 1982, 11(5): 341−356.

[40] Hu QH, Yu DR, Xie ZX. Numerical attribute reduction based on neighborhood granulation and rough approximation. Ruan Jian Xue Bao/Journal of Software, 2008, 19(3): 640−649 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/640.htm [doi: 10.3724/SP.J.1001.2008.00640]

[41] Hu QH, Zhang L, Zhang D, *et al*. Measuring relevance between discrete and continuous features based on neighborhood mutual information. Expert Systems with Applications, 2011, 38(9): 10737−10750.

[42] Skowron A, Rauszer C. The discernibility matrices and functions in information systems. In: Słowiński R, ed. Proc. of the Intelligent Decision Support. Theory and Decision Library (Series D: System Theory, Knowledge Engineering and Problem Solving). Berlin: Springer, 1992.

[43] Lin YJ, Hu QH, Liu JH, *et al*. Multi-label feature selection based on neighborhood mutual information. Applied Soft Computing, 2015, 38: 244−256.

[44] Fan X, Chen HM. Stepwise optimized feature selection algorithm based on discernibility matrix and mRMR. Computer Science, 2020, 47(1): 87−95 (in Chinese with English abstract). [doi: 10.11896/jsjkx.181202320]

[45] Huang GB, Zhou HM, Ding XJ, *et al*. Extreme learning machine for regression and multiclass classification. IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(2): 513−529.

[46] Peng HC, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226−1238.

[47] Li JG, Pan ZN, Su L, *et al*. Feature selection method LLE score used for tumor gene expressive data. Journal of Beijing University of Technology, 2015, 41(8): 1145−1150 (in Chinese with English abstract). [doi: 10.11936/bjutxb2014120005]

[48] Hu L, Gao WF, Zhao K, *et al*. Feature selection considering two types of feature relevancy and feature interdependency. Expert Systems with Applications, 2018, 93: 423−434.

[49] Sun L, Wang J, Wei JM. AVC: Selecting discriminative features on basis of AUC by maximizing variable complementarity. BMC Bioinformatics, 2017, 18(3): 50.

[50] Xie JY, Wang MZ, Zhou Y, *et al*. Differential expression gene selection algorithms for unbalanced gene datasets. Chinese Journal of Computers, 2019, 42(6): 1232−1251 (in Chinese with English abstract). [doi: 0.11897/SP.J.1016.2019.01232]

[51] Sammut C, Webb GI. Encyclopedia of Machine Learning. New York: Springer Science & Business Media, 2011.

[52] Wang R, Tang K. Feature selection for maximizing the area under the ROC curve. In: Proc. of the 2009 IEEE Int'l Conf. on Data Mining Workshops. Washington, DC: IEEE Computer Society, 2009. 400−405.

[53] Swets JA. Measuring the accuracy of diagnostic systems. Science, 1998, 240(4587): 1285−1293.

[54] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Bioinformatics, 2003, 1396−1400.

[55] Xie JY, Gao HC, Xie WX, *et al*. Robust clustering by detecting density peaks and assigning points based on fuzzy weighted *K*-nearest neighbors. Information Sciences, 2016, 354: 19−40.

**附中文参考文献**:

[7] 谢娟英, 高红超. 基于统计相关性与 *K*-means 的区分基因子集选择算法. 软件学报, 2014, 25(9): 2050−2075. http://www.jos.org.cn/1000-9825/4644.htm [doi: 10.13328/j.cnki.jos.004644]

[8] 谢娟英, 周颖. 一种新聚类评价指标. 陕西师范大学学报(自然科学版), 2015, 43(6): 1−8. [doi: 10.15983/j.cnki.jsnu.2015.06.161]

[9] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述. 控制与决策, 2012, 27(2): 161−166, 192.

[30] 刘华文. 基于信息熵的特征选择算法研究[博士学位论文]. 长春: 吉林大学, 2010.

[33] 李郅琴, 杜建强, 聂斌, 等. 特征选择方法综述. 计算机工程与应用, 2019, 55(24): 10−19. [doi: 10.3778/j.issn.1002-8331.1909-0066]

[37] 谢娟英, 吴肇中, 郑清泉. 基于信息增益与皮尔森相关系数的 2D 自适应特征选择算法. 陕西师范大学学报(自然科学版), 2020, 48(6): 69−81. [doi:10.15983/j.cnki.jsnu.2020.01.019]

[38] 毛峰. 大学物理: 下册. 武汉: 华中科技大学出版社, 2010.

[40] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报, 2008, 19(3): 640−649. http://www.jos.org.cn/1000-9825/19/168.htm [doi: 10.3724/SP.J.1001.2008.00640]

[44] 樊鑫, 陈红梅. 基于差别矩阵和 mRMR 的分步优化特征选择算法. 计算机科学, 2020, 47(1): 87−95. [doi: 10.11896/jsjkx.181202320]

[47] 李建更, 逄泽楠, 苏磊, 等. 肿瘤基因选择方法 LLE Score. 北京工业大学学报, 2015, 41(8): 1145−1150. [doi: 10.11936/bjutxb2014120005]

[50] 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. 计算机学报, 2019, 42(6): 1232−1251. [doi: 0.11897/SP.J.1016.2019.01232]

**谢娟英**(1971－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 数据挖掘, 生物医学数据分析.

**吴肇中**(1995－), 男, 硕士生, 主要研究领域为机器学习, 生物医学数据分析.