软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(3):1128−1140 [doi: 10.13328/j.cnki.jos.006099]

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

融合信息增益比和遗传算法的混合式特征选择算法*

许召召, 申德荣, 聂铁铮, 寇 月

(东北大学 计算机科学与工程学院, 辽宁 沈阳 110819)

通信作者: 许召召, E-mail: zhaozhaotoms@foxmail.com

摘 要: 随着信息技术以及电子病历和病案在医疗机构的应用, 医院数据库产生了大量的医学数据. 决策树因其分类精度高、计算速度快, 且分类规则简单、易于理解, 而被广泛应用于医学数据分析中. 然而, 医学数据固有的高维特征空间和高度特征冗余等特点, 使得传统的决策树在医学数据上的分类精度并不理想. 基于此, 提出了一种融合信息增益比排序分组和分组进化遗传算法的混合式特征选择算法(GRRGA). 该算法首先使用基于信息增益比的过滤式算法对原始特征集合进行排序, 然后按照密度等分的原理对排序后的特征进行分组, 最后再使用分组进化遗传算法对排序后的特征组进行遗传搜索. 其中, 分组进化遗传算法共分为种群内和种群外两种进化方法, 并使用两种不同的适应度函数来控制进化过程. 此外, 针对决策树的不稳定性, 提出使用 Bagging 方法对 C4.5 算法进行集成学习. 实验结果显示, GRRGA 算法在 6 组 UCI 数据集上的 Precision 指标均值为 87.13%, 显著优于传统的特征选择算法. 此外, 与另外两种分类算法对比可知, GRRGA 算法的特征筛选性能依然是最优的. 更重要的是, Bagging 方法在 Arrhythmia 和 Cancer 医学数据集上的 Precision 指标分别为 84.7%和 78.7%, 充分证明了该算法的实际应用意义.

关键词: 医学数据; 决策树; 特征选择; 遗传算法; 信息增益比

中图法分类号: TP18

中文引用格式: 许召召, 申德荣, 聂铁铮, 寇月. 融合信息增益比和遗传算法的混合式特征选择算法. 软件学报, 2022, 33(3): 1128−1140. http://www.jos.org.cn/1000-9825/6099.htm

英文引用格式: Xu ZZ, Shen DR, Nie TZ, Kou Y. Hybrid feature selection algorithm combining information gain ratio and genetic algorithm. Ruan Jian Xue Bao/Journal of Software, 2022, 33(3): 1128−1140 (in Chinese). http://www.jos.org.cn/1000-9825/6099.htm

Hybrid Feature Selection Algorithm Combining Information Gain Ratio and Genetic Algorithm

XU Zhao-Zhao, SHEN De-Rong, NIE Tie-Zheng, KOU Yue

(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract: In recent years, the application of information technology and electronic medical records and medical records in medical institutions has become more and more widespread, which has resulted in a large amount of medical data in hospital databases. Decision tree is widely used in medical data analysis because of its high classification precision, fast calculation speed, and simple and easily understood classification rules. However, due to the inherent high dimensional feature space and high feature redundancy of medical data, the classification precision of traditional decision trees is low. Based on this, this paper proposes a hybrid feature selection algorithm (GRRGA) that combines information gain ratio ranking grouping and group evolution genetic algorithm. Firstly, the information gain ratio based filtering algorithm is used to sort the original feature set; then, the ranked features are grouped according to the density principle of equal division; finally, a group evolution genetic algorithm is used to perform a search on the ranked feature groups. There are two kinds of evolution methods: in-population and out-population, which use two different fitness functions to control the evolution

* 基金项目: 国家自然科学基金(62172082，62072084, 62072086); 国家重点研发计划(2018YFB1003404)
收稿时间: 2020-01-23; 修改时间: 2020-03-09; 采用时间: 2020-04-09

process in group evolution genetic algorithm. The experimental results show that the average precision index of the GRRGA algorithm on the six UCI datasets is 87.13%, which is significantly better than the traditional feature selection algorithm. In addition, compared with the other two classification algorithms, the feature selection performance of the GRRGA algorithm proposed in this study is optimal. More importantly, the precision index of the bagging method on the arrhythmia and cancer medical datasets is 84.7% and 78.7% respectively, which fully proves the practical application significance of the proposed algorithm.

Key words: medical data; decision tree; feature selection; genetic algorithm; information gain ratio

随着信息技术以及电子病历和病案在医疗机构的广泛应用, 使得医院数据库产生了大量的医学数据[1,2]. 这些医学数据对于疾病的诊断、治疗和医学研究都是十分有价值的[3]. 决策树作为一种经典的分类算法, 因其分类精度高、运算速度快, 尤其是分类规则简单、易于理解, 而被广泛应用于医学数据分析中[4]. 然而, 由于医学数据所固有的高维特征空间和高度特征冗余等特点[5,6], 使得决策树在医学数据上的分类效果并不理想[7].

特征选择[8,9]是指从一组原始特征集合中筛选出最优的特征子集, 以达到降低特征维度和提升分类精度的目的. 根据是否与分类算法[10]有关, 可将特征选择算法分为过滤式(filter)、封装式(wrapper)和混合式(hybrid)这 3 种. 信息增益比[11]是一种经典的过滤式算法, 特征的信息增益比越大, 其包含的信息量也越大. 因此, 通过设置阈值来筛选信息增益比高于阈值的特征子集. 然而, 基于信息增益比的过滤式算法得到的特征子集辨识度较差[12,13]. 此外, 过滤式算法的阈值设定也具有一定的盲目性[14].

封装式算法[15]需要嵌入分类算法, 通过对原始特征集合进行搜索, 并使用分类算法作为评估算法. 遗传算法[16]是一种常见的搜索算法, 适合对较大特征空间数据进行搜索, 不需要评估算法具有单调性, 因此常被用于高维特征空间数据的特征选择中[17]. 但有研究表明[18]: 遗传算法容易陷入局部最优解, 出现“早熟”现象. 引起这种现象的主要原因是: 在进化初期, 群体中个体的多样性迅速降低, 使得算法过早收敛, 从而可能丢失一些有意义的搜索点和最优点, 而陷入局部最优.

由于过滤式算法和封装式算法各有优缺点, 因此有学者[19]提出将两种算法融合使用, 不仅可以节约时间, 而且提升了候选特征子集的辨识度. 混合式算法一般由两个阶段组成: 首先, 使用基于信息增益比的过滤式算法[11]剔除大部分冗余特征, 只保留少量特征, 从而有效地缩减了后续启发式算法的搜索规模; 然后, 将过滤式筛选的特征子集连同样本作为输入参数传递给基于遗传搜索的封装式算法[17], 以进一步筛选更重要的特征. 虽然混合式算法可以有效地克服过滤式和封装式的缺点. 然而, 过滤式阶段的阈值设定以及遗传算法的局部最优解等问题仍然需要解决[12−14,18].

基于上述描述, 本文提出一种融合信息增益比排序分组和分组进化遗传算法的混合式特征选择算法(GRRGA). 该算法主要分为 3 部分: 首先, 使用基于信息增益比的过滤式算法计算原始特征的信息增益比, 按照信息增益比的大小对原始特征进行排序; 然后对排序后的特征进行分组, 按照密度等分的方法对排序后的特征进行分组; 最后, 再使用基于分组进化遗传的封装式算法对排序后的特征组进行搜索, 按照种群内和种群外两种评估方法来筛选特征, 使用 C4.5 算法作为封装式算法中的评估算法. 此外, 本文使用集成方法对 C4.5 算法进行集成学习, 以降低 C4.5 算法的不稳定性. 在实验方面: 首先是本文所提 GRRGA 算法的验证实验, 通过在 6 组 UCI 数据集上将 GRRGA 算法与传统的特征选择算法进行对比实验, 以验证 GRRGA 算法的优越性; 然后, 使用集成方法对 C4.5 算法进行集成学习, 选择最优的集成方法作为应用实验的应用算法; 最后, 将本文所提 GRRGA 算法应用于 2 组 UCI 医学数据集上, 取得了实际的应用意义.

1 相关工作

特征选择是一种数据降维算法, 通过对高维特征空间数据进行搜索, 并使用评估算法对搜索的子集进行评估, 从而筛选出最优的特征子集. 过滤式和封装式的区别在于评估方法的不同[20], 而混合式则融合了两者的优点. 分别总结如下.

过滤式算法通过某种评估算法来增强特征与类别的相关性, 削减特征与特征之间的相关性. 常见的过

滤式算法有对称不确定性(symmetrical uncertainty, SU)[21]、信息熵(information entropy, IE)[22]和特征权重(ReliefF, RF)[23]等. Sosa 等人[21]利用互信息原理计算每个特征的对称性, 兼顾了特征与类别以及特征之间的相关性, 通过设置阈值来筛选特征. Mendoza 等人[23]提出了一种基于 ReliefF 的分布式过滤式算法, 该算法将 ReliefF 与 Spark 融合使用, 根据特征对近距离样本的区分能力进行特征评估, 用于解决庞大数据的特征选择问题. 也有学者[11]提出一种基于信息增益比(information gain ratio, GR)的过滤式算法: 首先, 使用信息增益比对特征进行排序; 然后, 根据特征之间的互信息进行特征选择.

以上过滤式算法中, 都是使用评估方法评价特征的重要性, 筛选的特征子集辨识度低, 且阈值的设定具有一定的盲目性. 因此, 有学者提出使用分类算法的分类精度作为特征子集的评估准则, 并结合搜索算法进行搜索[24]. 如 Zhang 等人[25]提出一种基于粒子群(partical swarm optimization, PSO)的封装式算法, 该算法使用粒子群算法对冗余特征进行筛选, 并使用 C4.5 算法作为评估算法. 也有学者[26]提出使用支持向量机来进行特征选择的 Wrapper 算法, 文中用遗传算法(genetic algorithm, GA)来寻找使得支持向量机的分类错误率最小的一组特征子集. 此外, 也有学者[15]提出一种改进遗传算法的封装式算法进行特征选择. 其中, 在遗传算法的种群进化中增加种群灭绝和转移策略, 以增强遗传算法进化过程中的种群多样性.

基于搜索策略的封装式算法中, 作者都是直接通过分类器的分类精度来评估特征的重要性[27]. 这类方法实际上是 Wrapper 算法和搜索策略的结合, 其结果较好, 但是运算的时间较长[28]. 此外, 遗传算法在进行特征子集搜索时容易出现“早熟”问题. 因此, 有学者提出将过滤式和封装式融合使用的混合式算法. 如 Uğuz 等人[29]提出一种混合式算法, 该算法首先使用信息增益方法对文档中的每个特征进行特征排序, 并设置阈值进行筛选; 然后, 再使用遗传算法和主成分分析融合方法对排序后的特征子集进行精选.

为了解决遗传算法的“早熟”问题, Ghareb 等人[30]提出一种增强型遗传算法和 6 种 Filter 算法融合的混合式算法: 首先, 使用信息增益和卡方检验等过滤式算法对特征集合进行初步筛选; 然后, 再使用增强型遗传算法对过滤后的特征子集进行选择. Lu 等人[19]提出一种最大化互信息和自适应遗传算法融合的混合式算法, 以过滤高维基因数据中的冗余特征. 其中, 最大化互信息用于特征集合的初步筛选, 自适应遗传算法和 4 个不同的分类器融合的 Wrapper 算法用于特征的进一步选择. Rani 等人[31]提出一种互信息和遗传算法结合的混合式算法, 用于高维基因数据的特征选择: 首先, 使用互信息筛选出高水平的特征子集; 然后, 再使用基于遗传算法的 Wrapper 算法进行选择. 其中, 使用支持向量机作为分类算法. 结果表明, 混合式算法选择的特征子集具有更高的分类精度.

在以上混合式算法中, 作者仅使用过滤式算法进行特征初选; 然后再使用封装式算法进行特征精选. 虽然融合了过滤式和封装式的优点, 但仍然面临以下问题: 1) 在过滤式中, 特征初选的阈值设定往往凭借经验, 具有一定盲目性; 2) 在封装式中, 传统的搜索策略存在一定的问题, 如遗传算法的“早熟”现象;3) 混合式算法仅仅是简单的融合, 并没有对过滤式和封装式进行深度的融合.

2 混合式特征选择算法

2.1 GRRGA: 混合式特征选择算法

传统的特征选择算法存在诸多问题, 如过滤式中阈值设定的盲目性问题以及封装式中遗传算法的“早熟”现象. 更重要的是, 单纯地将过滤式和封装式融合使用, 并没有充分利用两者的优点. 基于此, 本文提出一种融合信息增益比排序分组和分组进化遗传算法的混合式特征选择算法(GRRGA). GRRGA 算法的流程如图 1 所示.

由图 1 可知, GRRGA 算法主要分为 3 部分.

- 首先使用基于信息增益比 Filter 算法对原始特征进行排序;
- 然后根据密度等分的原理对排序后的特征进行分组, 将所有特征分为 k 个密度相等的特征组;
- 最后使用基于分组进化遗传算法的 Wrapper 算法对密度等分的特征组进行搜索, 并使用 C4.5 算法作为封装式算法的评估算法.

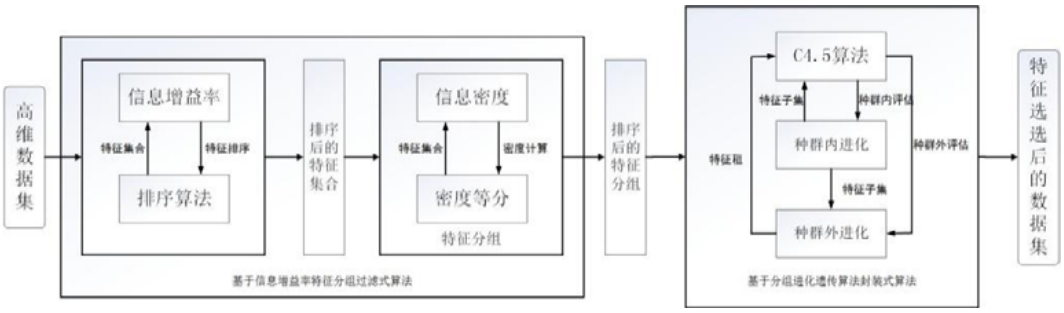


图 1 GRRGA 算法的流程图

2.1.1 特征排序

过滤式是进行封装式特征选择前的一种十分有效的特征初选方法, 通过设置阈值来筛选出更重要的特征子集, 在一定程度上消除特征冗余. 然而该方法具有一定盲目性, 阈值设定多凭经验, 过滤后的特征子集并不适用后续分类算法.

基于上述描述, 本文提出一种基于信息增益比排序分组的算法, 该算法使用信息增益比对初始特征集合进行排序, 以便调整和控制后续种群初始化在特征集合中的分布情况. 此外, 该算法按照各个特征所包含的信息增益比进行特征分组, 使得每个分组内的特征集合具有相同的信息密度. 首先给出一些关于信息密度的定义.

假设样本集 $S=\{x_1,x_2,...,x_n\}$, 非类别特征为 $F=\{F_1,F_2,...,F_{m-1}\}^T$, 样本集 S 的第 i 个特征为 F_i $(1\leqslant i\leqslant m-1)$, 第 j 个特征为 F_j $(1\leqslant j\leqslant m-1)$, 其中, $j<i$. 为了便于计算, 需要对所有特征进行归一化操作. 那么, 特征 F_i 的信息密度可以定义为

$$\rho_i = GR(F_i)\Big/\sum_{l=1}^{m-1} GR(F_i),\ \ i=1,2,...,m-1 \tag{1}$$

其中, $GR(F_i)$为第 i 个特征的信息增益比值. 通过公式(1), 可以得到两个特征之间的信息密度距离. 那么, 特征 F_i 和 F_j 之间的信息密度距离可以定义为

$$Dis_{ij} = \frac{GR(F_i)-GR(F_j)}{\sum_{l=1}^{m-1} GR(F_l)},\ \ i,j=1,2,...,m-1 \tag{2}$$

由公式(1)和公式(2)可知, $GR(F_i)$越大, 密度ρ_i越大, 那么特征 F_i 和 F_j 之间的信息密度距离越近, 其局部邻域内特征分布越稠密, 表明特征 F_i 和 F_j 在一个分组内的可能性越大, 因而特征间密度越大越好. 特征分组的目的是将权重密度高的组合在一块, 使得组内特征密度紧促, 组间密度疏松. 最后, 按照组内最大特征权重的方法对特征组进行重新排序, 使得有代表性的特征组排在最前. 算法描述如下.

算法 1. 基于信息增益比的特征排序分组算法.

输入: 原始特征集合 $F=\{F_1,F_2,...,F_{m-1}\}^T$; 特征分组块数为 k.

输出: 排序后的特征组 $\boldsymbol{GR}=\{\boldsymbol{GR}_1>\boldsymbol{GR}_2>\ldots>\boldsymbol{GR}_k\}$, $i=1,2,...,k$.

1) 初始化排序特征集合 $GR_F=\{\}$
2) **For** $i=1,2,...,m-1$ **do** \\对于所有非类别特征
3) 根据公式(7)–公式(10)计算特征 F_i 的信息增益比
4) 将计算信息增益比后的特征 F_i 加入集合 $GR_F=\{F_j+\ldots\}$
5) 输出排序后的特征集合 $GR=\{GR(F_1),GR(F_2),...,GR(F_{m-1})\}$
6) **End For**
7) **For** $i=1,2,...,m-1$ **do** \\对于排序后的所有非类别特征
8) **If** $GR(F_i)<GR(F_{i+1})$ **then**
9) $Swap(GR(F_i),GR(F_{i+1}))$ \\交换两个特征的顺序
10) **End If**

11) **End For**

12) **For** i=1,2,...,k **do**

13) **If** $GR(F_i) < GR(\sum_{l=1}^{m-1} F_l)/k$ **then**

14) 分别根据公式(5)和公式(6)计算特征的信息密度和密度距离

15) 将特征 F_i 加入第 i 个特征分组中 $\boldsymbol{GR}_i$

16) **End If**

17) **End For**

2.1.2 特征编码

特征编码: 在封装式特征选择过程中, 特征只有被选中或未被选中这两种情况, 故采用二进制编码方法. 与传统编码方法不同的是, 本文按照排序分组后的特征组进行编码, 即对排序后的特征组 $\boldsymbol{GR}$ 中的每个特征组进行二进制编码, 得到一个码长为 $k\times||GR||$的二进制串: $h_i=\{h_1,h_2,...,h_k\}$, i=1,2,...,k. h_i 为所有个体的组合, 称为个体的空间, 那么个体空间的大小为

$$C_{n_i}^1 + C_{n_i}^2 + ... + C_{n_i}^{n_i},\ i=1,2,...,k \tag{3}$$

种群初始化: 种群规模一般为个体编码长度的一个线性倍数.在实际应用中, 种群规模 m 一般取为 m–1 和 2(m–1)之间的一个确定数. 初始种群的选取一般使用随机选取的方法, 即使用随机函数产生 m–1 个基因组成的初始群体.

2.1.3 评估过程

封装式特征选择过程中的搜索往往使用分类算法对样本集进行建模训练, 然后根据分类结果作为特征集的适应度函数. 然而, 遗传算法在搜索过程中容易陷入局部最优解, 即在进化过程中, 种群个体丧失了种群多样性. 为了解决遗传算法的缺陷, 本文根据排序分组后的特征组, 提出两种种群评估方法, 通过种群内和种群外对初始化的种群进行评估.

对于种群内个体 $h_j=\{h_1,h_2,...,h_{n_i}\}, j=1,2,...,n_i; i=1,2,...,k$, 遗传算法的搜索过程中按个体的适应度进行选择, 适应度好的个体被遗传下一代的可能性大. 此外, 考虑到特征组内个体数目较少, 本文采用 C4.5 算法的分类精度作为个体评估的适应度函数, 定义为

$$f(h_j)=\frac{1}{M}\sum_{l=1}^{M}\gamma_l,\ l=1,2,...,M \tag{4}$$

其中, h_j 为第 n_i 组内的第 j 个个体, γ_l 表示 M 类分类问题中的每一类的分类精度.

对于种群外个体 $h_i=\{h_1,h_2,...,h_k\}$, i=1,2,...,k, 由于个体规模较组内个体增加了 1 倍, 考虑到特征选择的目的是选择尽可能少的特征子集, 故可对包含较多特征的个体给予一定的惩罚, 抑制这些个体选择概率而让包含较少特征的个体有更多的繁殖机会, 本文采用一种改进的适应度函数:

$$f(h_i)=\frac{1}{M}\sum_{l=1}^{M}\gamma_l-\lambda\frac{\sum_{j=1}^{m_i}\overline{h_i}}{\sum_{i=1}^{k}\sum_{j=1}^{m_i}h_i},\ j=1,2,...,m_i;\ i=1,2,...,k \tag{5}$$

其中, h_i 为组内筛选后的个体子集; λ为惩罚因子, 用于平衡分类精度和个体子集规模.

2.1.4 进化过程

遗传算法的进化操作主要包括选择操作、交叉操作和变异操作. 为了使适应度最优的个体遗传下去, 本文选择轮盘赌策略进行选择操作. 其具体过程为: 假设第 i 个个体的规模为 $n_i=||GR_i||$, i=1,2,…,k, 则 $h_j=\{h_1,h_2,...,h_{n_i}\}$, j=1,2,…,n_i; i=1,2,…,k 表示 p 个体的适应度为 $F(h_j)$, 那么它被选择的概率为

$$P(h_j)=\frac{F(h_j)}{\sum_{j=1}^{n_i}F(h_j)},\ j=1,2,...,n_i;\ i=1,2,...,k \tag{6}$$

由公式(6)可知, 个体的适应度越大, 其被选择的概率也就越大, 有利于算法的快速收敛. 此外, 使用选择操作选择的个体用于进化下去, 其中, 父代个体在下一代中存在的期望数目为 $G(h)=n_i\times P(h_j)$.

对于交叉和变异两种操作, 分别采用单点交叉和位变异策略. 对于单点交叉: 随机选择两个父代个体, 依据交叉概率 P_c 选择需要交叉的基因位置, 进而形成两个新的个体. 对于位变异: 对于两个父代个体, 依据变异概率 P_m 选择需要变异的基因位置, 然后按照 0 变 1 或 1 变 0 的方法进行变异, 从而形成新的个体(见算法 2).

算法 2. 分组进化遗传算法.

输入: 排序分组后的特征组 ***GR***.

输出: 最优特征子集.

初始化种群: P→随机产生 N 个长度为 1 的个体

1) **For** i=1,2,...,k **do**
2) **While** Gen≤Max_{Gen} 或 P_m≤0.09 **do** \\进化到最大代数或变异率小于 0.09 时终止进化
3) 种群内个体适应度评估: 根据公式(4)对 h_j 中的每个个体 h_{ni}, 计算 $f(h_j)$
4) 根据 $f(h_j)$产生新一代的种群 P_S
5) **If** $f(h_i)$=$f(h_{best})$ \\最优个体的适应度超过 3 代没有增加
6) P_m=P_m+0.02
7) 对个体进行选择、交叉和变异操作
8) 更新种群 P←P_s
9) **else**
10) P_m=0.033 \\置变异率
11) 对个体进行选择、交叉和变异操作
12) 更新种群: P←P_s
13) **End If**
14) **End While**
15) 种群外个体适应度评估: 根据公式(5)对 h 中的每组个体 h_i, 计算 $f(h_i)$
16) 将最优的 $f(h_i)$保存至数组 E_{best}[·]
17) **End For**
18) 从数组中返回最优特征子集

2.2 C4.5算法

决策树[32]是一种根据实例归纳的分类算法, 其模型构建过程可以分为训练集的训练过程和未知样本集的预测过程. 在使用决策树对未知样本集进行预测时, 该测试样本会从根节点沿着相应分支寻找属于该样本的类别. 因此, 利用决策树进行预测的关键是根据训练集构建决策树分类模型.

由于本文选取信息增益比作为混合式算法中的过滤式阶段, 而 C4.5 算法[33]根据信息增益比进行模型构建. 因此, 本文选取 C4.5 算法作为封装式的评估算法. 已知样本集 S=$\{x_1,x_2,...,x_n\}$, 每个样本 x_i 包含 m 个特征 F=$\{F_1,F_2,...,F_{m-1},F_m\}^T$. 假设类别特征 F_m 具有 k 个不同的取值, 那么根据 F_m 将样本集 S 划分为 k 个样本子集 C_k. 因此, 可以计算样本集 S 对分类的平均信息量为

$$I(S)=-\sum_{i=1}^{k}P(C_q)\log_2 P(C_q) \tag{7}$$

其中, $P(C_q)$=$|S_i|$/$|S|$, $|S_i|$和$|S|$分别为 S_i 和 S 样本的数目. C4.5 算法的构建过程就是使得划分后特征包含的信息量更加稳健的过程. 那么, 以离散特征 F_i (1≤i≤m–1)为例, 可以进行如下划分.

假设离散特征 F_i 具有 v 个不同的取值 F_i (1≤i≤v), 那么根据 F_i 的取值, 可以将样本集 S 划分为 v 个样本子集 $S_1,S_2,...,S_v$, 可以进一步将 $C_1,C_2,...,C_k$ 划分为 k×t 个子集, 每个子集 C_{pq} 表示 F_i 在 i=v 情况下, 属于第 p 类的样本集合. 那么, 对离散特征 F_i 进行划分后, 样本集 S 对分类的平均信息量为

$$I(S/F_i)=-\sum_{i=1}^{v}P(C_q)[-\sum_{p=1}^{k}P(C_{pq})\log_2 P(C_{pq})] \tag{8}$$

其中, $P(C_q)=\sum_{p=1}^{k}|C_{pq}|/|S|$, $P(C_{pq})=|C_{pq}|/|S|$. 根据非类别特征 F_i 进行划分后的样本信息量和条件信息量, 可以求得样本集 S 信息增益量 $G(S,F_i)$, 即

$$G(S,F_i)=I(S)-I(S,F_i) \tag{9}$$

由于使用特征 F_i 对 S 进行划分的信息增益比等于信息增益量与分裂信息量(split information, Sp)之比, 那么可以得到:

$$GR(S,F_i)=G(S,F_i)/Sp(S,F_i) \tag{10}$$

其中, 分裂信息量 $Sp(S,F_i)=-\sum_{l=1}^{t}(|S_l|/|S|)\log_2(|S_l|/|S|)$.

对于连续特征的样本集, C4.5 算法增加了离散化措施[34]. C4.5 算法的模型构建过程就是选择样本集 S 中具有最大信息增益比的特征作为分裂特征, 自上而下的方式完成决策树的构建过程. 然而, 有研究表明[35]: 决策树在特征分裂过程中存在较高的方差, 数据集的微小波动就会产生不同的分裂结果, 这就造成了决策树的不稳定性. 因此, 本文选取集成方法对 C4.5 算法进行集成学习, 以降低 C4.5 算法的不稳定性.

3 实验结果与分析

3.1 实验设置

为了验证本文所提 GRRGA 算法的性能, 我们选取了 6 组 UCI 普通数据集和 2 组 UCI 医学数据集进行实验. 其中, 6 组 UCI 普通数据集分别来自不同的领域, 如 Spambase 主要用于垃圾邮件的识别分类, Comma 主要用于监控物流运输情况, Scadi 主要用于测试残疾儿童的自我护理情况等. 2 组 UCI 医学数据集分别为 Arrhythmia 和 cancer. 其中, Arrhythmia 为心率失常数据集, 该数据集的样本数量为 452 个, 共包含年龄、性别、身高和心率等 279 个特征, 主要用于区分患者是否存在心律不齐的情况; 而 Cancer 为肺癌数据集, 该数据集的样本数量为 32 个, 共包含 56 个特征, 主要用于区分患者是否患有癌症. 所有数据的详细介绍见表 1.

表 1 实验数据集描述

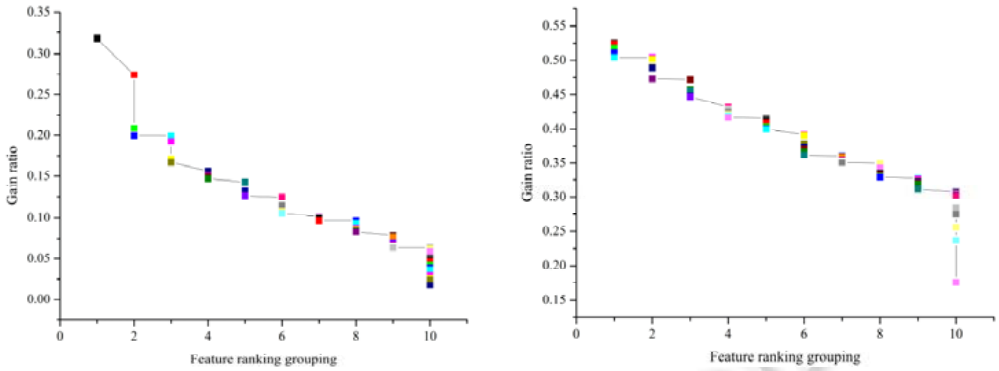
数据集属性	数据集名称	数目	特征	类别
验证数据	eighthr	4 736	72	2
	Spambase	4 601	57	2
	Comma	3 942	96	3
	Meu	2 856	71	56
	Urban	675	147	9
	Scadi	70	205	7
医学数据	Arrhythmia	452	279	2
	Cancer	32	56	3

特征选择算法通过使用分类器的分类精度来评估筛选特征子集的好坏. 因此, 本文在实验中分别对原始数据集和经过特征筛选后的数据集, 使用分类器的分类精度进行评估. 实验共分为验证实验和应用实验. 在验证实验中, 首先是本文所提 GRRGA 算法的特征选择实验, 通过设置中断来观察 GRRGA 算法的过滤式阶段和封装式阶段; 然后, 将 GRRGA 算法与传统的特征选择算法进行对比验证; 最后, 将封装式阶段中的 C4.5 算法与另外 4 种分类算法[36]进行对比验证, 并使用两种集成方法[37]对 3 种决策树算法进行集成学习, 以降低决策树的不稳定性. 与验证实验一样, 使用相同的算法和参数进行应用实验.

3.2 验证实验

3.2.1 GRRGA 算法的特征筛选实验

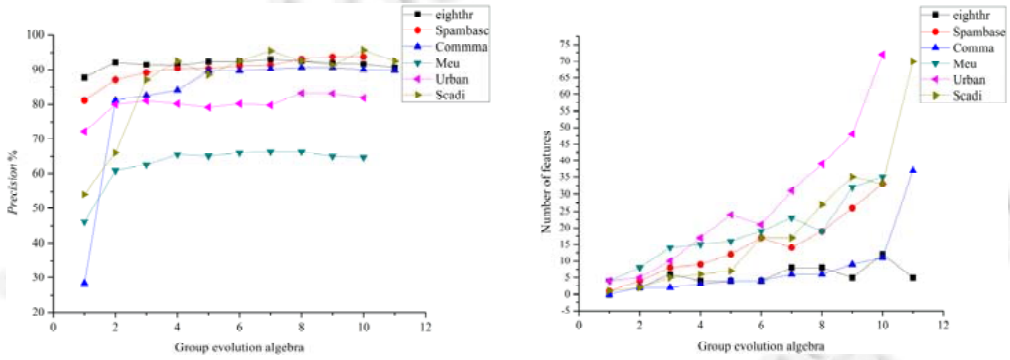
本节是所提 GRRGA 算法的特征选择实验. 根据 GRRGA 算法的工作原理, 实验可分为 3 部分: 首先是特征排序过程, 即计算数据集中每个特征的信息增益比, 按照信息增益比的大小对特征进行排序; 然后, 使用密度等分的方法对排序后的特征进行分组; 最后是分组进化遗传算法的搜索过程, 并使用 C4.5 算法作为评估算法进行特征选择. 图 2 给出了 Spambase 和 Meu 数据集特征排序后的分组图.



(a) Spambase 数据集 (b) Meu 数据集

图 2 数据集的特征排序后的分组图

由图 2 可知, 根据特征的信息增益比的大小对原始特征进行排序, y 轴的每个分组代表该组内的特征子集, 如 Spambase 数据集上 Feature ranking grouping=10 时, 表示第 10 个特征组有 14 个特征. 观察图 2(a)和图 2(b), 特征的信息增益比越小, 意味着越多的特征被分为一组. 信息增益比越大, 意味着可能只有一个特征被分为一组(如 Spambase 数据集上 Feature ranking grouping=1). 另外, 信息增益比为 0 的特征并未被分组, 但该类特征依然被用于种群初始化过程中. 那么, 分组进化遗传算法的种群外进化过程可由图 3 给出.



(a) GRRGA 算法筛选的分类精度 (b) GRRGA 算法筛选的特征子集规模

图 3 分组进化遗传算法的中群外进化过程

由图 3(a)可知, 随着进化代数的递增, C4.5 算法在 6 组 UCI 数据集上的 Precision 指标呈现先升后降的趋势. 当进化代数开始时, C4.5 算法的 Precision 指标均达到最低值. 这主要是由于种群的个体过少导致遗传算法在进化过程中出现“早熟”现象, 从而陷入局部最优解. 观察图 3(b)发现, 随着进化代数的递增, 特征子集的数目呈现上升趋势, 在个别数据集上出现下降情况, 但是整体仍然为上升. 因此, 本文综合考虑了 C4.5 算法的 Precision 和特征子集的数目, 如 Urban 数据集选取第 8 代为最优种群规模.

3.2.2 与传统特征选择算法的对比实验

为了对比 GRRGA 算法的优越性, 本节选取了 10 种传统特征选择算法进行对比实验. 其中, 过滤式算法分别为 GR[11]、ReliefF[23]和 SU[21], 封装式算法分别为 GA[26]、PSO[25]和 EA[24], 混合式算法分别为 GRGA、GREA 和 GRPSO. 为了最大限度地体现 GRRGA 算法在特征选择过程中的筛选能力, 在传统混合式算法的过滤式阶段使用相同的阈值参数. 对于 3 种传统过滤式算法的阈值根据经验进行设置. 使用 C4.5 算法对特征选择后的数据集进行测试. 实验结果见表 2 和表 3.

表 2 GRRGA 算法与传统特征选择算法在 6 组 UCI 数据集上的分类精度

数据集	原始	Filter			Wrapper			Hybrid			
		GR	ReliefF	SY	GA	EA	PSO	GRGA	GREA	GRPSO	GRRGA
eighthr	92.1	91.4	91.1	91.4	92.2	88.2	92.2	91.9	91.7	92.2	**93.1**
Spambase	93.0	90.9	91.0	92.7	93.1	93.0	93.5	91.3	91.1	90.7	**93.8**
Comma	89.5	88.8	88.4	88.7	**90.9**	90.2	90.4	90.0	90.6	90.2	90.6
Meu	64.2	65.2	42.8	64.7	65.2	64.8	65.5	65.5	65.8	**66.4**	**66.4**
Urban	80.2	82.1	77.6	77.9	79.4	82.4	81.9	82.6	82.6	**83.7**	83.2
Scadi	88.8	90.0	90.0	90.0	92.6	92.6	90.8	**95.7**	92.6	91.4	**95.7**
平均	84.63	84.73	80.15	84.23	85.57	85.20	85.72	86.17	85.73	85.77	**87.13**

表 3 GRRGA 算法与传统特征选择算法在 6 组 UCI 数据集上的候选特征子集规模

Dataset	原始	Filter			Wrapper			Hybrid			
		GR	ReliefF	SY	GA	EA	PSO	GRGA	GREA	GRPSO	GRRGA
eighthr	72	37	39	40	4	8	3	8	10	4	8
Spambase	57	24	25	18	29	30	36	19	17	16	26
Comma	96	41	64	41	51	37	37	11	14	18	6
Meu	71	28	42	30	39	32	35	22	21	19	19
Urban	147	64	35	35	79	68	65	30	32	22	39
Scadi	205	141	124	141	42	69	36	33	34	31	33
平均	108.00	55.83	54.83	50.83	40.67	40.67	35.33	20.50	21.33	18.33	21.83

由表 2 和表 3 可知, C4.5 算法在 6 组原始数据集上的 Precision 指标均值仅为 84.63%. 这主要是由于数据的高维特征空间和高度特征冗余极大地损害了 C4.5 算法的分类性能. 观察 3 种传统的过滤式算法发现, 在不同数据集上呈现不同的结果, 但整体而言, 经过 GR 算法处理后的效果更优. 对比 3 种传统的封装式算法可知, 使用 PSO 作为搜索策略的效果最优, C4.5 算法在 6 组数据集上的 Precision 均值均为 85.72%, 结果略有优势. 无论是过滤式还是封装式算法, 均存在经过特征选择后的特征子集辨识度变差的情况.

对比 3 种传统的混合式算法可知, 在不同数据集上呈现不同的结果: 在数据集 Scadi 上, GRGA 算法的结果最优; 而在数据集 Meu 上, 算法 GRPSO 的结果更好. 但整体而言, GRGA 算法在 6 组数据集上的 Precision 指标均值为 86.17%, 稍微优于另外两种算法. 对比本文所提 GRRGA 算法和传统的混合式算法可知, 在 Comma 和 Urban 数据集上的结果并不是最优的. 但整体而言, 本文所提 GRRGA 算法在 6 组 UCI 数据集上的 Precision 指标均值为 87.13%.

3.2.3 与其他分类算法的对比实验

由第 3.3.2 节实验可知, 本文所提 GRRGA 算法在 6 组 UCI 数据集上的特征选择效果是最优的. 但是, 封装式阶段使用 C4.5 算法作为评估算法具有一定的局限性. 因此, 本节提出使用另外两种分类算法(NB 和 KNN)进行对比实验, 以验证本文所提 GRRGA 算法的性能的适用性. 此外, 为了使实验更具公正性, 本文还选取了传统的 GRGA 算法作为对比算法. 实验结果如表 4 和表 5 所示.

由表 4 可知, 与 C4.5 算法一样, NB 和 KNN 在原始数据集上的分类精度都是较差的. 因此, 数据的高维特征空间极大地损害了算法的分类性能. 3 种分类算法在不同的数据集上具有不同的结果, 如: 在 Spambase 等数据集上, C4.5 算法的分类精度优于其他算法; 而在 eighthr 数据集上, KNN 算法的结果更优. 但整体而言, C4.5 算法在 6 组 UCI 数据集上的 Precision 指标均值为 84.63%, 优于 NB 和 KNN.

表 4 3 种分类算法在 6 组 UCI 数据集上的分类精度

数据集	C4.5			NB			KNN		
	原始	GRGA	GRRGA	原始	GRGA	GRRGA	原始	GRGA	GRRGA
eighthr	92.1	91.9	**93.1**	**93.1**	87.8	87.8	92.1	**93.1**	92.5
Spambase	93.0	91.3	**93.8**	84.2	89.3	89.7	90.8	90.4	92.8
Comma	89.5	90.0	90.6	80.3	95.8	**98.7**	75.8	89.0	90.5
Meu	64.2	65.5	66.4	74.1	65.4	**76.1**	46.0	57.1	59.2
Urban	80.2	82.6	83.2	82.1	84.8	**86.3**	76.1	84.8	84.2
Scadi	88.8	95.7	95.7	86.2	90.2	**96.9**	85.7	93.6	**96.9**
平均	84.63	86.17	87.13	83.33	85.55	**89.25**	77.75	84.67	86.01

表 5 3 种分类算法在 6 组 UCI 数据集上的候选特征子集规模

数据集	原始	C4.5		NB		KNN	
		GRGA	GRRGA	GRGA	GRRGA	GRGA	GRRGA
eighthr	72	8	8	1	1	20	25
Spambase	57	19	26	14	8	16	25
Comma	96	11	6	7	1	3	2
Meu	71	22	19	15	44	15	21
Urban	147	30	39	21	46	32	9
Scadi	205	33	33	40	11	44	6
平均	108.00	20.50	21.83	16.33	18.50	21.67	14.67

对比本文所提 GRRGA 算法与传统的 GRGA 算法可知, GRRGA 算法不仅有效地降低了数据的特征数目, 而且 3 种分类算法在 GRRGA 算法筛选后的数据集上的分类精度有了显著的提升, 在 6 组 UCI 数据集上的 Precision 指标均值分别为 87.13%, 89.25%和 86.01%, 较原始数据集上分别提升了 2.50%, 5.92%和 8.26%.此外, 经过 GRRGA 算法筛选后的数据集的特征子集规模是最少的, 显著缩短了 C4.5 算法的建模预测时间.

3.3 应用实验

3.3.1 与传统特征选择算法的对比实验

本节给出了传统的过滤式、封装式和混合式算法与本文所提 GRRGA 算法在 2 组 UCI 医学数据集上的对比实验. 与验证实验一样, 应用实验中使用 C4.5 算法作为测试算法, 实验参数与验证实验中的参数一样. 表 6 和表 7 给出了 C4.5 算法在特征选择后的医学数据集上的实验结果.

表 6 GRRGA 算法与传统特征选择算法在 2 组 UCI 医学数据集上的分类精度

数据集	原始	Filter			Wrapper			Hybrid			
		GR	ReliefF	SY	GA	EA	PSO	GRGA	GREA	GRPSO	GRRGA
Arrhythmia	79.3	78.8	78.4	78.6	80.7	76.3	78.2	80.1	77.6	78.3	**83.4**
Cancer	40.6	68.4	55.2	71.9	63.3	80.1	63.8	**77.4**	75.7	75.7	**77.4**
平均	59.95	73.60	66.80	75.25	72.00	78.20	71.00	78.75	76.65	77.00	**80.40**

表 7 GRRGA 算法与传统特征选择算法在 2 组 UCI 医学数据集上的候选特征子集规模

数据集	原始	Filter			Wrapper			Hybrid			
		GR	ReliefF	SY	GA	EA	PSO	GRGA	GREA	GRPSO	GRRGA
Arrhythmia	279	131	114	131	120	140	90	55	65	44	26
Cancer	56	8	31	8	19	16	23	3	4	4	3
平均	167.50	69.50	72.50	69.50	69.50	78.00	56.50	29.00	34.50	24.00	14.50

由表 6 可知, 与验证实验的结果一样, C4.5 算法在 2 组 UCI 医学数据集上的分类结果都是较差的. 观察传统的过滤式算法发现, C4.5 算法在 3 种过滤式算法筛选后的数据集上的 Precision 指标都是极差的. 这主要是由于医学数据集含有更多的冗余特征. 同样地, 在传统的封装式算法和混合式算法中具有相同的结果. 对比传统的混合式算法和本文所提 GRRGA 算法可知, GRRGA 算法在 2 组 UCI 医学数据集上的 Precision 指标分别为 83.4%和 77.4%, 在所有特征选择算法中的结果是最优的. 不仅如此, 由表 7 可知, GRRGA 算法筛选的特征子集规模也是最小的. GRRGA 算法以最小的代价, 获取最好的效果.

3.3.2 与其他分类算法的对比实验

由验证实验可知, C4.5 算法在高维特征空间数据集上的分类效果最优. 但是在医学数据集上是否具有同样的结果还需进一步验证. 与验证实验一样, 使用 GRGA 算法作为对比算法, 并使用相同的参数进行实验. 表 8 和表 9 给出了 3 种分类算法在 2 组 UCI 医学数据集上的实验结果.

表 8 3 种分类算法在 2 组 UCI 医学数据集上的分类精度

数据集	C4.5			NB			KNN		
	原始	GRGA	GRRGA	原始	GRGA	GRRGA	原始	GRGA	GRRGA
Arrhythmia	79.3	80.1	**83.4**	78.4	82.7	82.4	64.2	76.9	79.0
Cancer	40.6	77.4	77.4	63.1	81.5	**84.7**	50.7	82.5	82.5
平均	59.95	78.75	80.40	70.75	82.10	**83.55**	57.45	79.70	80.75

表 9　3 种分类算法在 2 组 UCI 医学数据集上的候选特征子集规模

数据集	原始	C4.5		NB		KNN	
		GRGA	GRRGA	GRGA	GRRGA	GRGA	GRRGA
Arrhythmia	279	55	26	63	50	30	39
Cancer	56	3	3	5	22	3	3
平均	167.50	29.00	14.50	34.00	36.00	16.50	21.00

由表 8 和表 9 可知, 3 种分类算法在原始 UCI 医学数据集上的分类精度都是极差的. 观察 GRRGA 与传统 GRGA 发现, 经过 GRRGA 算法筛选后的特征子集具有更优的分类精度, C4.5、NB 和 KNN 算法的 Precision 指标均值分别提升了 20.45%, 12.80%和 23.30%, 而且选取 C4.5 算法作为 GRRGA 的评估算法所筛选的特征子集最小(14.50). 针对决策树的不稳地性, 本文选取了两种集成方法(Bagging 和 Adaboost)对 3 种决策树算法进行集成学习, 表 10 和表 11 给出了两种集成方法的实验结果.

表 10　Bagging 在 2 组 UCI 医学数据集上的分类精度

数据集	C4.5			Randomtree			REPtree		
	原始	GRGA	GRRGA	原始	GRGA	GRRGA	原始	GRGA	GRRGA
Arrhythmia	83.2	83.8	**84.7**	76.2	80.7	82.4	80.9	82.3	83.0
Cancer	68.9	78.7	**78.7**	52.3	78.5	78.5	65.6	70.9	73.7
平均	76.05	81.25	**81.70**	64.25	79.60	80.45	73.25	76.60	78.35

表 11　Adaboost 在 2 组 UCI 医学数据集上的分类精度

数据集	C4.5			Randomtree			REPtree		
	原始	GRGA	GRRGA	原始	GRGA	GRRGA	原始	GRGA	GRRGA
Arrhythmia	80.6	**84.5**	81.2	68.4	71.4	72.5	76.5	78.5	78.5
Cancer	45.9	77.4	77.4	42.7	**78.7**	**78.7**	50.2	75.9	75.9
平均	63.25	**80.95**	79.30	55.55	75.05	75.60	63.35	77.20	77.20

由表 10 和表 11 可知, 两种集成方法在 2 组原始 UCI 医学数据集上的集成效果是十分显著的. 与未经集成学习的结果一样, C4.5 是 3 种决策树算法中的分类效果最优的, 而且经过 GRRGA 筛选后的特征子集具有更高的分类精度. 对比两种集成方法发现, Bagging 在 3 种决策树算法中的集成效果最优的, Precision 指标均值分别为 81.70%, 80.45%和 78.35%, 优于 Adaboost 对 3 种决策树算法的集成效果. 因此, 选取 Bagging 作为 C4.5 算法的集成方法能够实现精准的临床决策.

4　结　论

针对医学数据固有的高维特征空间和高度特征冗余, 本文提出了一种融合信息增益比特征排序和分组进化遗传算法的混合式特征选择算法(GRRGA). 该算法主要分为 3 部分: 首先, 使用信息增益比计算每个特征的信息增益比值, 与传统的过滤式算法不同的是, 本文仅对特征进行排序; 然后, 根据特征的信息密度进行密度等分, 不同密度的特征组具有不同的特征数量; 最后, 使用分组进化遗传算法对密度等分的特征组进行搜索, 并使用 C4.5 算法的分类精度进行评估. 其中, 评估算法分为种群内和种群外两种评估策略. 在实验方面, 首先在 6 组 UCI 数据集上进行验证实验, 结果显示 C4.5 算法在 GRRGA 算法特征筛选后的数据集上的 Precision 指标均值为 87.13%, 显著优于其他特征选择算法. 此外, 通过与另外两种决策树算法对比, 得出本文所提 GRRGA 算法的适用性; 最后, 在 2 组 UCI 医学数据集的应用实验结果显示, C4.5 算法在 GRRGA 算法特征筛选后的数据集上的 Precision 指标均值为 80.40%, 显著优于传统特征选择算法; 而且选取 Bagging 作为集成方法, 显著地提升了 C4.5 算法的分类精度. 尽管本文所提 GRRGA 算法较传统的特征选择算法在高维数据集上的效果有了较好的提升, 但在高维小样本数据中的效果并不理想(如 Sacdi 数据集), 后续我们将重点研究高维小样本数据中的特征选择问题.

References:

[1] Koleck TA, Dreisbach C, Bourne PE, *et al*. Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. Journal of the American Medical Informatics Association, 2019, 26(4): 364–379.

[2] Ghazavi SN, Liao TW. Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine, 2008, 43(3): 195−206.

[3] Chen J, Li K, Rong H, *et al*. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Information Sciences, 2018, 435: 124−149.

[4] Gao W, Bao W, Zhou X. Analysis of cough detection index based on decision tree and support vector machine. Journal of Combinatorial Optimization, 2019, 37(1): 375−384.

[5] Pölsterl S, Conjeti S, Navab N, *et al*. Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection. Artificial intelligence in medicine, 2016, 72: 1−11.

[6] Liu CZ, Wang YJ. Research on medical data mining and its applications. Journal of Biomedical Engineering, 2014, 31(5): 1182−1186 (in Chinese with English abstract).

[7] Park CH, Kim SB. Sequential random *k*-nearest neighbor feature selection for high-dimensional data. Expert Systems with Applications, 2015, 42(5): 2336−2342.

[8] Liu Y, Cao JJ, Diao XC, Zhou X. Survey on stability of feature selection. Ruan Jian Xue Bao/Journal of Software, 2018, 29(9): 2559−2579 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5394.htm [doi: 10.13328/j.cnki.jos.005394]

[9] Sayed GI, Hassanien AE, Azar AT. Feature selection via a novel chaotic crow search algorithm. Neural Computing and Applications, 2019, 31(1): 171−188.

[10] Abedinia O, Amjady N, Zareipour H. A new feature selection technique for load and price forecast of electrical power systems. IEEE Trans. on Power Systems, 2016, 32(1): 62−74.

[11] Karegowda AG, Manjunath AS, Jayaram MA. Comparative study of attribute selection using gain ratio and correlation based feature selection. Int'l Journal of Information Technology and Knowledge Management, 2010, 2(2): 271−277.

[12] Hsu HH, Hsieh CW, Lu MD. Hybrid feature selection by combining filters and wrappers. Expert Systems with Applications, 2011, 38(7): 8144−8150.

[13] Chen Y, Cheng XQ, Li Y, Dai L. Lightweight Intrusion detection system based on feature selection. Ruan Jian Xue Bao/Journal of Software, 2007, 18(7): 1639−1651 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/1639.htm [doi: 10.1360/jos181639]

[14] Hancer E, Xue B, Zhang M. Differential evolution for filter feature selection based on information theory and feature ranking. Knowledge-based Systems, 2018, 140: 103−119.

[15] Xue X, Yao M, Wu Z. A novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm. Knowledge and Information Systems, 2018, 57(2): 389−412.

[16] Zhang L, Zhang B. Research on the mechanism of genetic algorithms. Ruan Jian Xue Bao/Journal of Software, 2000, 11(7): 945−952 (in Chinese with English abstract). http://jos.org.cn/jos/article/abstract/20000712?st=article_issue

[17] Das AK, Das S, Ghosh A. Ensemble feature selection using bi-objective genetic algorithm. Knowledge-based Systems, 2017, 123: 116−127.

[18] Eroglu DY, Kilic K. A novel hybrid genetic local search algorithm for feature selection and weighting with an application in strategic decision making in innovation management. Information Sciences, 2017, 405: 18−32.

[19] Lu H, Chen J, Yan K, *et al*. A hybrid feature selection algorithm for gene expression data classification. Neurocomputing, 2017, 256: 56−62.

[20] Hu Z, Bao Y, Xiong T, *et al*. Hybrid filter-wrapper feature selection for short-term load forecasting. Engineering Applications of Artificial Intelligence, 2015, 40: 17−27.

[21] Sosa-Cabrera G, García-Torres M, Gómez-Guerrero S, *et al*. A multivariate approach to the symmetrical uncertainty measure: Application to feature selection problem. Information Sciences, 2019, 494: 1−20.

[22] Zhang X, Mei C, Chen D, *et al*. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. Pattern Recognition, 2016, 56: 1−15.

[23] Palma-Mendoza RJ, Rodriguez D, De-Marcos L. Distributed ReliefF-based feature selection in spark. Knowledge and Information Systems, 2018, 57(1): 1−20.

[24] Xue B, Zhang M, Browne WN, *et al*. A survey on evolutionary computation approaches to feature selection. IEEE Trans. on Evolutionary Computation, 2015, 20(4): 606−626.

[25] Zhang Y, Wang S, Phillips P, *et al*. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-based Systems, 2014, 64: 22−31.

[26] Welikala RA, Fraz MM, Dehmeshki J, *et al*. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. Computerized Medical Imaging and Graphics, 2015, 43: 64−77.

[27] Wang L, Liu SJ, Chen BL, *et al*. Heuristic discrimination cotton ripeness using hybrid filter and wrapper. Journal of Computer Research and Development, 2013, 50(2): 269−277 (in Chinese with English abstract).

[28] Lee SJ, Xu Z, Li T, *et al*. A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making. Journal of Biomedical Informatics, 2018, 78: 144−155.

[29] Uğuz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-based Systems, 2011, 24(7): 1024−1032.

[30] Ghareb AS, Bakar AA, Hamdan AR. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications, 2016, 49: 31−47.

[31] Rani MJ, Devaraj D. Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification. Journal of Medical Systems, 2019, 43(8): 1−11.

[32] Tanha J, van Someren M, Afsarmanesh H. Semi-supervised self-training for decision tree classifiers. Int'l Journal of Machine Learning and Cybernetics, 2017, 8(1): 355−370.

[33] Xu P, Lin S. Internet traffic classification using C4.5 decision tree. Ruan Jian Xue Bao/Journal of Software, 2009, 20(10): 2692−2704 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3444.htm [doi: 10.3724/SP.J.1001.2009.03444]

[34] He M, Zhang J, Vittal V. Robust online dynamic security assessment using adaptive ensemble decision-tree learning. IEEE Trans. on Power Systems, 2013, 28(4): 4089−4098.

[35] Han L, Li W, Su Z. An assertive reasoning method for emergency response management based on knowledge elements C4.5 decision tree. Expert Systems with Applications, 2019, 122: 65−74.

[36] Lohita K, Sree A A, Poojitha D, *et al*. Performance analysis of various data mining techniques in the prediction of heart disease. Indian Journal of Science and Technology, 2015, 8(35): 1−7.

[37] Gomes HM, Barddal JP, Enembreck F, *et al*. A survey on ensemble learning for data stream classification. ACM Computing Surveys (CSUR), 2017, 50(2): 1−36.

附中文参考文献:

[6] 刘婵桢, 王友俊. 医学数据挖掘技术与应用研究. 生物医学工程学杂志, 2014, 31(5): 1182−1186.

[8] 刘艺, 曹建军, 刁兴春, 周兴. 特征选择稳定性研究综述. 软件学报, 2018, 29(9): 2559−2579. http://www.jos.org.cn/1000-9825/5394.htm [doi: 10.13328/j.cnki.jos. 005394]

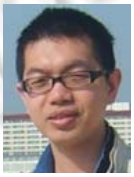
[13] 陈友, 程学旗, 李洋, 戴磊. 基于特征选择的轻量级入侵检测系统. 软件学报, 2007, 18(7): 1639−1651. http://www.jos.org.cn/1000-9825/18/1639.htm [doi: 10.1360/jos181639]

[16] 张铃, 张钹. 遗传算法机理的研究. 软件学报, 2000, 11(7): 945−952. http://jos.org.cn/jos/article/abstract/20000712?st=article_issue

[27] 王玲, 刘善军, 陈兵林, 等. 混合过滤器和封装器启发式判别籽棉成熟度. 计算机研究与发展, 2013, 50(2): 269−277.

[33] 徐鹏, 林森. 基于 C4.5 决策树的流量分类方法. 软件学报, 2009, 20(10): 2692−2704. http://www.jos.org.cn/1000-9825/3444.htm [doi: 10.3724/SP.J.1001.2009.03444]

许召召(1991－), 男, 博士生, CCF 学生会员, 主要研究领域为数据挖掘, 机器学习.



聂铁铮(1980－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为数据质量, 数据集成.

申德荣(1964－), 女, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为分布式数据管理, 数据集成.

寇月(1980－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为实体搜索, 数据挖掘.