信息与控制　2019 年　第 48 卷　第 2 期：224 ~ 231

DOI：10.13976/j.cnki.xk.2019.8283　　文章编号：1002 - 0411(2019) - 02 - 0224 - 08

决策信息系统的连续型特征选取方法

李国和[1,2]，杨绍伟[1,2]，吴卫江[1,2]，郑艺峰[1,2,3]

1. 中国石油大学(北京)石油数据挖掘北京市重点实验室，北京　102249；
2. 中国石油大学(北京)地球物理与信息工程学院，北京　102249；
3. 闽南师范大学计算机学院数据科学与智能应用福建省高等学校重点实验室，福建 漳州　363000

通信作者：杨绍伟，787866446@qq.com　收稿/录用/修回：2018 - 06 - 06/2018 - 09 - 05/2018 - 10 - 29

摘要

在大数据应用过程中，对特征集合进行约简，降低数据维度，有助于提升数据模型的泛化能力. 采用随机森林模型选择和相似性度量结合的方式对特征集合进行特征初选，并通过前向搜索策略以距离为评价方式对初选集合进行二次筛选，最终获得特征子集. 算法模型采用局部遍历以提高执行效率，同时通过前向选择算法解决传统方法无法确定最优特征数目的问题. 实验结果表明，本文提出的方法能更有效地选择特征子集，提高模型的分类准确率.

关键词

特征约简
随机森林
相似性度量
二次筛选

中图法分类号：TP18

文献标识码：A

A Continuous Feature Selection Method of Decision Information System

LI Guohe[1,2], YANG Shaowei[1,2], WU Weijiang[1,2], ZHENG Yifeng[1,2,3]

1. *Beijing Key Lab of Petroleum Data Mining*, *China University of Petroleum*, *Beijing* 102249, *China*;
2. *College of Geophysics and Information Engineering*, *China University of Petroleum*, *Beijing* 102249, *China*;
3. *Key Laboratory of Data Science and Intelligence Application* (*Fujian Province*), *School of Computer Sciences*, *Minnan Normal University*, *Zhangzhou* 363000, *China*

Abstract

In the process of large data application, it is necessary to reduce the feature set for improving the generalization ability of the data model. We use random forest model selection and similarity measure to select feature sets. Then, we adopt the forward search strategy to finish the second filtering. In the algorithmic model, it uses local traversal because it can be helpful to enhance the execution efficiency. At the same time, it can effectively solve the problem about how to determine the optimal number of features. The experimental results show that this method can obtain the feature subset more effectively and improve the classification accuracy.

Keywords

feature reduction;
random-forest;
similarity measure;
two filtering

0　引言

随着智能技术的发展，信息技术已进入大数据时代，为了更好地获得潜在的有价值的数据信息，需要对数据进行预处理. 特征选择是数据预处理的重要环节，能有效地提高模型的泛化能力，减少过拟合的情况，增强系统的可解释性.

特征选择方式包括：过滤式(filter)，封装式(wrapper)和嵌入式(embedded)三种方式[1]. 在连续性决策系统的特征选择过程中，一般采用嵌入式方法和封装式方法，而过滤式方法通常用在封装式方法的评价函数之中.

嵌入式方法主要通过构建模型对特征进行排序，运行速度快，但其精度主要依赖于模型对数据的处理能力. Breiman 等人提出一种基于 l1 正则化的支持向量机模型，主要使用分类准确率对特征进行评价[2]；常春云等人使用 Lasso 方法进行特征排序，并使用支持向量机模型进行自闭症预测[3]；赵宇等人提出基于集成特征选择的最优化支持向量机模型，通过参数调节和控制分类精度和特征数目，分别求解最优分类和最少特征[4]；傅昊等人提出随机森林和递归特征消除相结合的方式，通过随机重采样技术和节点随机分裂技术，构建多棵回归树，最终采用扰动特征取值或移除节点的方式来判定特征的重要程度，移除排名最后的特征，反复迭代多次，直至特征数目满足要求为止[5]. 上述方法，最优特征子集数目需人为确定，模型无法有效甄别相似特征，若存在相似性较高的特征，容易出现冗余.

封装式方法主要采用遗传算法与评价函数相结合的方式，提取特征子集. 封装式特征选取方法无需指定特征数目，其收敛速度较慢，最终结果过度依赖于初值的选择和参数的选取. Hancer 等人提出基于人工蜂群算法的封装式方法，采用支持向量机进行特征子集评价[6]；张文倩等人采用多群体遗传算法来处理基于分层梯度方向直方图的特征向量选择[7]；Mafarja 等人提出基于改进的鲸鱼优化算法，结合模拟退火算法，能有效提高执行效率，但复杂度较高[8].

嵌入式方法无法有效获得最优特征数目，并且存在大量冗余，同时特征子集选取结果依赖模型对该空间分布样本的处理能力[9]. 封装式方法则具有较高的时间复杂度，同时参数众多，初值选取依赖经验，最终选出的特征子集效果并不一定较优[10]. 为了更好地解决上述问题，在保证执行效率的基础上使特征子集更加精简、有效，本文将随机森林与相似性相结合，对特征集合进行初选，再采用前向搜索(SFS)的方式以距离为评价准则进行二次筛选，从而寻找性能更优的特征子集.

1　相关概念

1.1　连续型特征的决策信息系统

定义 1(决策表信息系统)　$K=\langle U, R, V, f\rangle$，对于 $\forall R_i \subseteq R$，构成 K 的特征子系统 $K_i=\langle U, R_i, V, f\rangle$. 其中，$U$ 表示论域，$U=\{\cup u_i \mid i=1, 2, \cdots, |U|\}$；$R=\{\cup r_i \mid i=1, 2, \cdots, |R|\}$，$R=C\cup D$($C$ 和 D 分别为条件属性集和决策属性集，在决策信息系统中 $D\neq\varnothing$)；$V=\{Uv_i \mid i=1, 2, \cdots, |R|\}$，表示所有特征取值的集合，$v_i$ 表示第 i 个特征的值域；$f=U\times R\rightarrow V$ 为信息函数，表示对象 $u_i\in U$ 在 r_i 上的投影.

定义 2(连续型特征的决策信息系统)　属性取值为连续型数据的决策表信息系统，表示为 $\forall c_i\in C$，$c_i\times R\rightarrow v_i$，满足 $v_{ij}-\varepsilon\in v_i$. 其中，$C$ 为条件属性集合，v_i 为条件属性 c_i 的值域，为某样本在第 i 个属性的取值，$v_{ij}-\varepsilon$ 代表该属性取值的邻域.

1.2　随机森林特征重要性度量

定义 3(随机森林)　由一组回归树分类器组成，$\{h(X, \boldsymbol{\theta}_k) \mid k=1, 2, \cdots, K\}$. 其中，$\boldsymbol{\theta}_k$ 表示服从独立同分布的随机向量，K 为随机森林中回归树的个数. 给定自变量 X，每个回归树分类器通过投票来决定最终的分类结果[11].

定义 4(随机森林算法的分类正确率)　分类结果正确占总样本的比例，即：

$$\text{Accuracy}=(\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \tag{1}$$

其中，TP(true positive)表示分类结果中正确的正例；TN(true negative)表示正确的负例；FP(false positive)表示分类错误的正例；FN(false negative)表示分类错误的负例.

定义 5(随机森林重要性度量)　对第 j 个特征重要性的度量标准为袋外数据自变量值发生轻微扰动后的分类正确率与扰动前分类正确率的平均减少量，即：

$$D_j=\frac{1}{n}\sum_{i=1}^{B}R_b-R_{bj} \tag{2}$$

其中，B 表示训练样本个数；R_{bj} 表示决策树 T_b 对袋外数据正确分类的个数；R_{bj} 表示对第 j 个特征进行随机扰动后模型的正确分类个数[12].

1.3　相关性度量

定义 6(皮尔逊相关系数)　两个变量之间的协方差和标准差的商，即：

$$\rho(\boldsymbol{x}, \boldsymbol{y})=\frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma_x\sigma_y}=\frac{E((\boldsymbol{x}-\mu_x)(\boldsymbol{y}-\mu_y))}{\sigma_x\sigma_y} \tag{3}$$

定义 7(相关系数)　使用联合概率密度函数描述网格化后向量相关程度，即：

$$\begin{aligned} I(\boldsymbol{x}, \boldsymbol{y}) &= \int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}p(\boldsymbol{x}, \boldsymbol{y})\operatorname{lb}\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \\ &\approx I(X, Y) \\ &= \sum_{X, Y}p(X, Y)\operatorname{lb}\frac{p(X, Y)}{p(X)p(Y)} \end{aligned} \tag{4}$$

其中，$\boldsymbol{x}$、$\boldsymbol{y}$ 表示向量；X、Y 表示通过给定参数 i 和 j 对 $\boldsymbol{x}-\boldsymbol{y}$ 联合分布进行 i 行 j 列网格化后的数据.

定义 8(最大信息数)　不同尺度下相关系数的最大值[13]，即在分布集中区域通过联合概率密度来计算的相关系数，计算方式如下：

$$\text{MIC}(\boldsymbol{x}, \boldsymbol{y})=\max_{|X||Y|<B}\frac{I(\boldsymbol{x}, \boldsymbol{y})}{\operatorname{lb}(\min(|X|, |Y|))} \tag{5}$$

其中，B 表示经验值，取值为数据总量的 0.6 或者 0.55 次方；$|X|$、$|Y|$ 表示网格化后 $\boldsymbol{x}$、$\boldsymbol{y}$ 取值的最大值，用以确定分布区域.

1.4　样本距离度量

定义 9(欧氏距离)　欧氏距离源自 N 维欧氏空间中两点 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$，那么 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 间的距离公式[14]：

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j)=\sum_{i=1}^{N}\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^2} \tag{6}$$

其中，$\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 皆为 N 维空间中的向量.

2　随机森林—序列前向选择

本文提出随机森林—序列前向选择(RSFS)算法，无需人为指定特征子集数目，能自动获得较优特征子集. 同时，算法的时间复杂度小于遗传算法. RSFS 算法采用特征子集中的特征与决策类别尽可能相关，特征子集中特征之间尽可能无关的原则，对特征子集进行约简.

RSFS 算法先对相邻特征列向量进行比较处理，剔除相似特征，再通过启发式搜索获得最终特征子集. 算法主要包括随机森林特征初选和启发式搜索两个阶段.

2.1　序列向前搜索

定义 10(特征选择)　对决策信息系统 $K=\langle U, R, V, f\rangle$ 中条件属性进行约简的过程，约简结果 $K_i=\langle U, R_i, V, f\rangle$. 其中，$K_i$ 表示特征子系统；$R_i\in R$ 且 $D_i=D$.

特选选择过程如图 1 所示，包括产生过程、评价函数、停止准则和验证过程四部分[15]，具体过程为：

1) 产生过程(generation procedure)：搜索特征子集，负责为评价函数提供特征子集.

2) 评价函数(evaluation function)：评价一个特征子集的好坏程度，用 GetScore(K_i)表示.

3）停止准则（stopping criterion）：设置评价函数的满足条件，若满足则迭代停止.

4）验证过程（validation procedure）：验证特征子集的有效性.

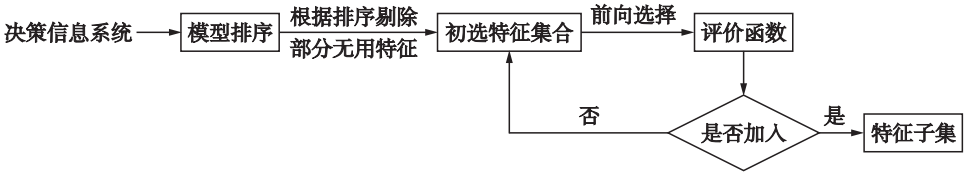

图 1 特征选择流程

Fig.1 The flow of feature selection

定义 11（序列向前搜索） 指在特征选择过程中，特征子集 C_i 从∅开始，每次选择一个特征 c_i 加入特征子集 C_i 构成特征子系统，使评价函数 GetScore（K_i）最优的过程，其中每次迭代加入一个特征.

2.2 随机森林特征初选

1）随机森林特征排序

特征初选采用过随机森林算法对特征进行度量，使用袋外数据分类准确率来评估特征的重要性. 将特征按照其重要程度由高到低排列，剔除重复的或意义不大的特征，提高启发式搜索二次筛选的效率，具体过程如算法 1 所示.

RF 中每颗 CART 回归树采取是均方差 mse 作为节点分裂标准. 为了保证每棵 CART 树结构相对简单，将每棵树的可选特征 m_{try} 设置为$\sqrt{N}$（N 为特征数目）[16].

算法 1 RandomForest

RandomForest（K, T_num = 100, $k=\sqrt{|C|}$）
//T_num-决策树数目，k-每棵树选取特征数目
//$K=\langle U, R, V, f\rangle$

将样本划分为训练样本和袋外数据，袋外数据大小设置为 $0.2\times|U|$ //样本划分
$D'=[0, \cdots, 0]$ //初始化$|C|$个特征重要性为全0
For($i=1$; $i<$nums; i++) //设置决策树数目 nums
  n=CART(rand(k)) //随机选取 k 个特征构造树
  For($j=0$; $j<k$; j++) //对树的特征进行扰动
    由式(1)，式(2)求出 D_j
    $D[j]'+=D_j$ //累加重要性
  Endfor //结束一棵树
EndFor //结束所有树
index=Getindex(order(D'))//得到按重要性升序排序后特征的下标
Return index //得到排序的特征索引 index

2）相似度方法

特征选择的过程推理表明，高度相似的特征更有可能作为相邻位置. 因此，本文进行特征向量相关程度对比时，仅仅考虑排名相邻两个特征的相似性，而进行全局比较，降低算法时间复杂度. 算法主要剔除相似性较高的特征列向量及离散程度较差的特征向量，具体过程如算法 2（FirstSelect）所示.

在比较列向量方差时，采用式（7）进行特征标准化处理[17].

$$\text{Norm}(\boldsymbol{x})=\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \tag{7}$$

其中，$\|\boldsymbol{x}\|_2$ 代表特征向量的 2 范数.

进行相似性度量时，采用式（8）对皮尔逊相关系数 Pearson 和最大信息系数进行描述.

$$\text{Sim}(\boldsymbol{x}_i, \boldsymbol{x}_j)=0.5\times\rho(\boldsymbol{x}_i, \boldsymbol{x}_j)+\text{MIC}(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{8}$$

算法 2 FirstSelect

FirstSelect(K, index,)
//index 为之前的排序索引集合，β 为方差阈值
//$K=\langle U, R, V, f\rangle$

For($i=|C|$; $i>2$; i--) //遍历除前两个之外的其它特征向量
  $\boldsymbol{x}_{\text{index}[i]}=\text{Norm}(\boldsymbol{x}_{\text{index}[i]})$ //对该特征向量进行标准化
    If(var($\boldsymbol{x}_{\text{index}[i]}$)$<\beta\times$var($\boldsymbol{x}_{\text{index}[i]}$)):
    index − =index[i] //若方差小于指定阈值，则丢弃该特征
  Else:
    For($j=1$; $j<3$; j++) //对之前两个特征向量相似性进行比较
      由式(7)得出相似性
        If(Sim($\boldsymbol{x}_{\text{index}[i]}$, $\boldsymbol{x}_{\text{index}[i-j]}$)
        >Sim($\boldsymbol{x}_{\text{index}[i]}$,$\boldsymbol{x}_{\text{index}[|\text{index}|-1]}$)):
        index − =index[i] //如果和之前特征相似，则删去该特征
        Endif
    Endfor //结束一个特征判断
  Endif
EndFor //结束全部特征的判断
Return index //得到排序的特征索引 index

2.3 前向选择

对于初选所获得特征集合，本文使用序列前向的启发式搜索方法对其进行二次处理，采取近邻点分类正确率作为评价标准进行一趟扫描，得出最终特征子集[18]，具体过程如算法 4（SecondSelect）所示.

本文主要参考 ReliefF 确定特征重要性的方式[19]，通过比较部分样本及其在空间中近邻样本类别是否相同，来判断特征有效性，具体评价方式如算法 3（GetScore）所示.

算法 3　GetScore

```
GetScore (K_i)
 //K_i = ⟨U, C_temp ∪ D, V, f⟩, 即特征子集 C_temp构成的决策表
U1 = Random(U, 100)          //从样本中有放回地随机选取 100 个
Score = 0
U1 = Scale(U1)               //标准化数据, 用于计算距离
For(i = 0; i < 100; i ++)    //遍历之后所有选出的样本
  Mindistance = + ∞          //初始化最近距离
  Index j = 0                //初始化最近点坐标
  For(j = 0; j < |U|; j ++)//遍历所有样本, 选出最近邻点
      If(j! = i and D_Ctemp(U_i, U_j) < Mindistance): //其中
    D_Ctemp为按式(6)计算的在当前特征子空间中的距离
          Mindistance = D_Ctemp(U_i, U_j)
          Index j = j      //选出当前子集空间下的最近点的索引
      Endif
  EndFor
  If(D[i] == D[index j]):          //对近邻点类别相同的样本进行统计
      Score ++
  Endif
EndFor
Return Score                 //得到最终分数
```

算法 3 返回结果为[0, 100]之间的数字，该数字表示当前特征空间中，选中样本集合近邻点与选中样本集类别相同的数目，以此作为对特征子集的评价值. 当样本数目较少时，又放回抽样保证了可以选出 100 个样本；当样本数目极多时，近邻点的选取可以无需遍历整个样本集合，而在选中的样本集合中进行，尽管准确率有所降低，但保证了较快的执行速度.

算法 4　SecondSelect

```
SecondSelect(K, index)
//index-之前的排序索引集合
//K = ⟨U, R, V, f⟩
C_temp = {index[0]}
Score = 0  //初始化特征子集, 将最重要的特征加入, 初始化评价函数
For(i = 1; i < |index|; i ++)    //遍历之后所有特征向量
  K_i = K[:C_temp]           //抽取特征子集对应的特征空间
  STMP = GetScore(K_i)  //保存加入特征前子集的评价值
  Ctemp + = index[i]         //加入 index 中第 i 个特征
  If(GetScore(K_i) ≤ STMP): //如果不能使系统评价值提升
      Ctemp + = index[i]           //除去该特征
  Endif
EndFor                             //结束全部特征的判断
Return Ctemp                 //得到最终特征子集
```

3　实验验证

本文实验主要采用两种方式，即同数目准确率比较，同准确率数目比较来验证 RSFS 算法有效性. 具体操作分别为使用 RSFS 方法和其它算法获取相同数目的特征子集，使用特征子集构造人工神经网络模型比较与其它方法的分类准确率；以及划定准确率标准，进行参数调整，得到满足准确率的子集为止，比较 RSFS 方法与其它方法特征子集的数目.

实验数据采用 7 个 UCI 数据集(如表 1 所示)，实验过程中，将试验数据集划分为训练数据和测试数据. 实验环境为 Windows10(64 位)，8 G 内存，Intel(R) Core(TM) 2 CPU E7300@2.66 GHz.

表 1　数据集信息

Tab.1　Information of data set

数据集(本实验中的编号)	样本数	特征数	类别数
Hill-Valley(1)	606	101	2
LSVT_voice_rehabilitation(2)	126	309	2
breast_cancer_wisconsin(3)	198	34	2
BlogFeedback(4)	60 021	281	8
gesture_phase_dataset(5)	9 900	50	5
Dataset for Sensorless Drive Diagnosis(6)	58 509	49	11
Connectionist Bench(7)	208	60	2

3.1　参数设置

本文实验对比 2 种过滤式方法、一种封装式方法和 6 种嵌入式方法，按照 4:1 比例划分训练集、测试集来构建人工神经网络[20].

对比方法具体为：过滤式方法相关系数法(Correlation)及最大信息系数法(MIC)法[21]，封装式方法贪心搜索的最大相关最小冗余法(MRMR)[22]、嵌入式方法线性回归法(Linear-Regression)、L1 正则化线性回归法(Lasso)，L2 正则化线性回归法(Ridge)[23]、随机森林回归法(RFR, random forest regressor)[24]、稳定的特征选择法(stability selection)[23]和基于 L2 正则化线性回归模型的递归特征消除法(RFE_lr)[25]. 具体参数设置见表 2.

表 2　方法参数说明

Tab.2　Parameter description of the method

方法名	参数名	参数取值
correlation	null	null
MIC	B	$0.5 \times \lvert X \rvert \lvert Y \rvert$
MRMR	null	null
LinearRegression	null	null
Lasso	α	0.05
Ridge	β	0.07
RFR	K, fea_num	100, $\sqrt{\lvert C \rvert}$
Stability	null	null
RFE_lr	β	7

1) 相关系数法及最大信息系数法无需参数；

2）最大相关最小冗余法（MRMR）利用启发式算法得出候选特征子集，相关性度量采用皮尔逊相关系数；

3）线性回归无需指定参数；

4）L1 正则化线性回归法，正则项参数取为 0.05；

5）L2 正则化线性回归法，正则项参数取为 7；

6）随机森林回归法使用回归树，树的数目 $K=100$，生成树随机抽取的特征数目为 $\sqrt{|C|}$，其中 C 为条件属性，回归树节点分裂采取是均方差 mse 作为标准，采取扰动特征的取值或移除节点的方式考查特征的重要程度；

7）稳定的特征选择法使用线性回归模型对特征排序；

8）递归特征消除法使用 L2 正则化线性回归模型进行特征评分，L2 正则项参数取为 7.

3.2 超参数调节

参数 β 的本质是对低方差特征进行过滤，针对本次试验所用数据集，通过网格参数寻优法，寻找合理的方差阈值 β，删除低方差特征. 实验采用不同参数的 RSFS 方法，选出特征子集. 在相同特征子集下，对比人工神经网络下的测试集分类错误率. 分类模型采用人工神经网络模型，以分类错误率来描述特征子集的性能. 人工神经网络参数设置为：采用 10 个隐层，每个隐层 30 个神经元；梯度下降时学习率为 1e-1；Batch_Size 设为 10；L2 正则项系数设置为 0.001. 具体寻优过程如表 3 所示[26].

表 3 不同 β 参数下各数据集的分类误差率

Tab.3 The classification error rate of each data set under different parameter β

β	1	2	3	4	5	6	7
0.1	0.25	0.24	0.10	0.23	0.20	0.20	0.19
0.2	0.18	0.23	0.11	0.18	0.16	0.16	0.16
0.3	0.09	0.08	0.05	0.07	0.16	0.15	0.06
0.4	0.05	0.05	0.10	0.01	0.02	0.05	0.05
0.5	0.08	0.07	0.01	0.04	0.09	0.07	0.01
0.6	0.07	0.11	0.01	0.10	0.09	0.09	0.00
0.7	0.14	0.10	0.06	0.10	0.09	0.08	0.02
0.8	0.14	0.13	0.11	0.09	0.10	0.09	0.08
0.9	0.21	0.21	0.24	0.22	0.19	0.18	0.18
1.0	0.24	0.31	0.29	0.32	0.22	0.19	0.17

3.3 准确性对比分析

相关系数、线性回归等对比方法无法直接找出最优特征，且取遍所有可能数目 1 到 N（其中 N 为特征总数）以寻找最优特征数目复杂度太高，不具备操作意义. 针对这种情况，本文进行了两组实验，即在相同特征数目下比较分类能力以及在相同分类能力下比较特征数目的多少.

首先，RSFS 算法使用最优参数 $\beta=0.4$ 得到特征子集，比较不同特征算法下选择相同数目特征分类误差率，如表 4 所示.

表 4 误差率及子集特征数

Tab.4 Error rate and subset feature number

数据集编号（特征数）	1(13)	2(14)	3(3)	4(11)	5(14)	6(17)	7(9)
原数据	0.27	0.24	0.06	0.37	0.18	0.27	0.27
Correlation	0.13	0.16	0.09	0.16	0.10	0.13	0.08
MIC	0.14	0.13	0.07	0.45	0.19	0.05	0.16
MRMR	0.12	0.09	0.07	0.10	0.19	0.10	0.14
LinearRegression	0.13	0.20	0.15	0.12	0.06	0.20	0.09
Lasso	0.09	0.22	0.05	0.14	0.23	0.19	0.17
Ridge	0.14	0.08	0.09	0.14	0.01	0.15	0.18
RFR	0.19	0.08	0.10	0.16	0.18	0.21	0.14
Stability	0.12	0.22	0.09	0.18	0.02	0.02	0.22
RFE_lr	0.12	0.20	0.03	0.10	0.08	0.08	0.09
RSFS	0.06	0.06	0.02	0.07	0.01	0.08	0.08

实验结果表明，相对于其它算法，本文算法在编号 1、2、3、4、5、7 的 6 个数据集中，去除超过 80% 的特征；在编号 6 的数据集中，去除接近 70% 的特征. 由此可见，RSFS 方法在保留样本信息的同时，能有效地降低样本维度，更好提升数据模型的泛化能力. 特征选取的意义在于避免过拟合，理论上往往针对分类精度要求越高、分类越多的数据越有效，由上图可以看出，多分类数据集中表现最好的数据集 4 中，相比属性规模相似的数据集 2 较其它的选择方法，高出了 1% 的正确率.

同时相比原数据本方法准确率有了大幅提升，在对分类精度要求较高的多分类问题中尤为明显，在数据集 3 ~ 数据集 5 中使用本方法提升了几乎 20% 的分类准确率.

从误差率对比折线图（如图 2 所示）中，结合表 2 信息，可以看到 RSFS 的预测能力优于其它模型. 在特征数目较多的二分类数据集中（1、2、3 号数据集），RSFS 方法误差率最低，比表现次好的 Lasso、Ridge、RFE_lr 方法分别高出了超过 3%、2% 及 0.5% 的正确率. 在同样特征数目同样较多的多分类数据集中（4、5 号数据集），效果同样明显，RSFS 比表现次好的 MRMR 和 RFR 方法分别高出了超过 3%、0.1% 的正确率. 在特征稍少的情况下，由于特征相似的情况较少，此时 RSFS 的准确率略低. 在二分类数据集中（7 号数据集），RSFS 只比表现最好的算法 Correlation 差 0.7% 的正确率. 在 6 号数据集中，RSFS 的误差率高于 Stability 方法 6%，MIC 方法 3% 排名第 3.

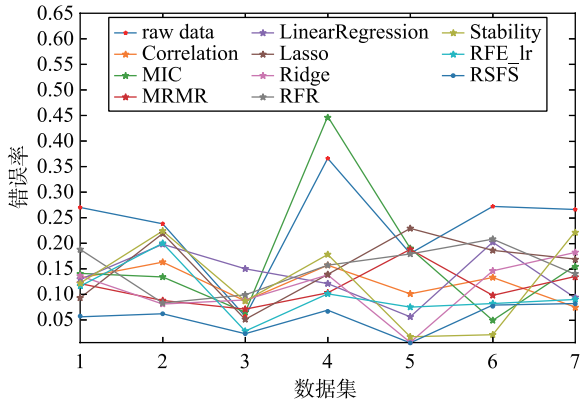

图 2　不同算法相同数目特征子集误差率

Fig.2　Error rate of the same number feature subset of different algorithms

之后，进行相同误差率下特征数目的比较，设置测试集误差率阈值 err 为 10%．调节 β 参数，7 个数据集参数分别为 0.4、0.5、0.5、0.5、0.4、0.4、0.3，使 RSFS 方法的 err 低于 10%．此时可以确定 RSFS 方法的特征数目．对于对比方法，将贪心搜索 MRMR 算法的停止条件设置为 err 低于 10%，可确定该方法特征数目；针对其它排序式特征选择算法，设置 $k=1$，求出 k 个特征对应的特征子集误差率，k 进行自加操作，直到 err 低于 10% 时停止自加，此时 k 为排序方法的特征数目．各种方法 err 低于 10% 的特征数目如表 5 所示．实验表明，在误差率相近的情况下，RSFS 方法能获得更小的特征子集．

表 5　err 低于 10% 的特征数

Tab.5　The number of feature when err is under 10%

数据集	1	2	3	4	5	6	7
Correlation	14	48	12	33	11	33	21
MIC	21	33	11	22	7	23	12
MRMR	32	17	7	26	15	16	8
LR	26	31	7	13	10	11	11
Lasso	35	46	12	15	12	11	16
Ridge	23	7	9	7	9	19	9
RFR	31	11	4	9	14	17	7
Stability	9	16	15	13	6	13	13
RFE_lr	11	29	11	16	11	18	12
RSFS	8	10	3	11	13	15	5

实验结果表明，RSFS 方法在考虑特征之间的相似性的同时，也充分考虑决策属性之间的相关性．RSFS 只需要调节相关参数，就能有效解决不同数据集下模型优化问题．

3.4　时间效率对比分析

针对算法性能比较，本文选择两种方式：进行一趟特征处理流程，得出相同数目特征子集的耗时和得到相似准确率特征子集的耗时．为使时间差距明显，选择样本和特征数目都较大的数据集 4 进行，RSFS 方法在前向搜索进行特征打分时，采用选定样本集合内取近邻点的策略．

首先，使用最优参数寻找 RSFS 方法的最优特征子集，将其它方法处理到相同数目(11)特征子集，比较计算时间．

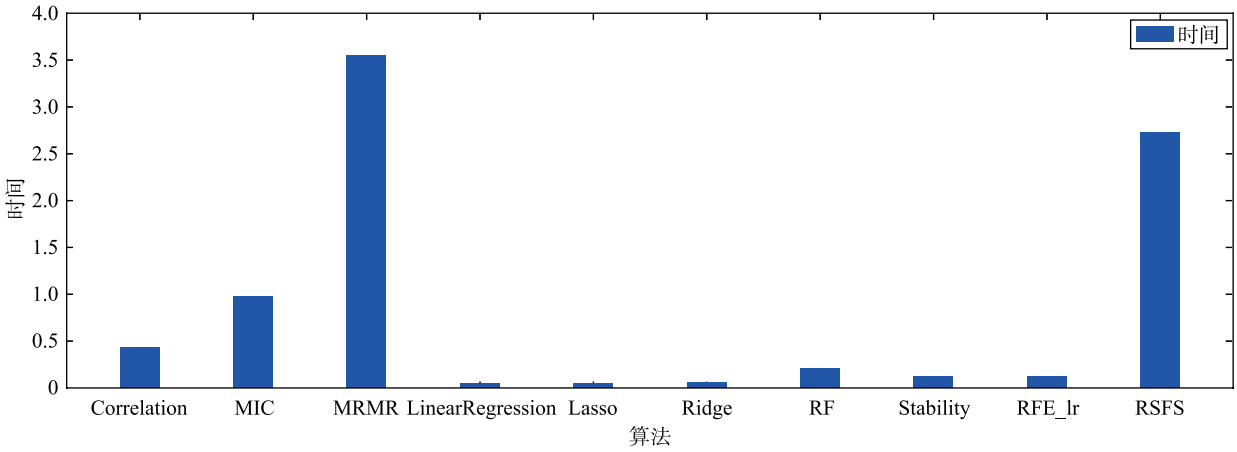

图 3　不同算法相同数目特征子集运行时间

Fig.3　Running time of the same number feature subset of different algorithms

由图 3 可以看到，尽管同数目有着更高的正确率，但是在运行效率上处理得到指定数目特征的问题 RSFS 方法并不占优．RSFS 方法的优势在于可以确定最优特征数目，即在保证正确率达标情况下具有更快的处理速度．

由此可见，单次处理时间基本为多次过滤式方法和一次嵌入式方法的加和，但在选择最优子集上，本方法更省时，同时更具有操作性．

4　结论

本文采用两次筛选的方式对特征进行约简，在考虑特征与决策属性相关性的同时，还兼顾特征之间的冗余性．在去除特征冗余性的过程中，采用局部遍历的方法，比较按重要性降序排列特征向量近邻之间的相似性，大大降低运行时间复杂度．同时，在启发式搜索的比较过程中，采

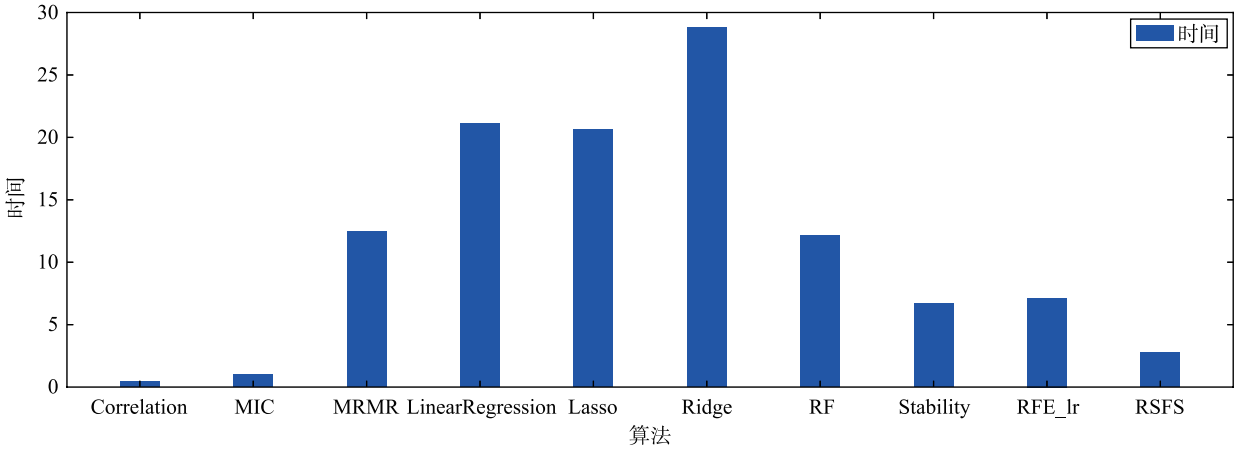

图 4 得到错误率低于 10% 的特征子集的运行时间

Fig.4 Running time of getting feature subset with error rate below 10%

用贪心搜索策略，且评价函数时间复杂度较低，一定程度上解决封装式方法运行速度较慢的问题.

另一方面，在二次搜索特征的过程中，利用启发式方法能自动获得特征子集，无需指定特征数目. 两次搜索过程考虑特征之间相似性，一定程度上解决了过滤式和嵌入式方法无法确定最优特征子集、模型难以处理特征冗余的问题. 本文结合两个方面的优点，可以得到更优的特征子集. 算法通过实现特征集合的自动选取，无需指定太多的参数，避免在数据预处理过程中投入大量精力，有利于分析高维数据，能在实际应用中取得更好的应用效果.

参考文献

[1] 黄铉. 特征选择研究综述[J]. 信息与电脑(理论版), 2017(24): 67 - 68.
Huang X. Summary of research on feature selection[J]. Information and Computer (Theoretical Version), 2017(24): 67 - 68.

[2] Breiman L. Rand om forests[J]. Machine Learning, 2001, 45(1): 5 - 32.

[3] 常春云. 基于 Lasso 特征选择的自闭症预测[J]. 北京生物医学工程, 2017, 36(6): 564 - 568, 596.
Chang C Y. Autism prediction based on Lasso feature selection[J]. Biomedical Engineering in Beijing, 2017, 36(6): 564 - 568, 596.

[4] 赵宇, 陈锐, 刘蔚. 集成特征选择的最优化支持向量机分类器模型研究[J]. 计算机科学, 2016, 43(8): 177 - 182, 215.
Zhao Y, Chen R, Liu W. Optimization of ensemble feature selection based on support vector machine classifier[J]. Computer Science, 2016, 43(8): 177 - 182, 215.

[5] 傅昊, 徐国胜. 基于随机森林和 RFE 的组合特征选择的研究[C]//第十九届全国青年通信学术年会论文集. 北京: 中国通信学会, 2014.
Fu H, Xu G S. Research on the combination of feature selection based on random forests and RFE[C]//Proceedings of the 19th National Youth Communication Annual Conference. Beijing: China Institute of Communications, 2014.

[6] Hancer E, Xue B, Zhang M, et al. Pareto front feature selection based on artificial bee colony optimization[J]. Information Sciences, 2017, 422: 462 - 479.

[7] 张文倩, 庄华亮, 陈翔, 等. 基于竞争思想的分级聚类算法[J]. 信息与控制, 2017, 46(5): 614 - 619.
Zhang W Q, Zhuang H L, Chen X, et al. Hierarchical clustering algorithm based on competitive thinking[J]. Information and Control, 2017, 46(5): 614 - 619.

[8] Mafarja M M, Mirjalili S. Hybrid whale optimization algorithm with simulated annealing for feature selection[J]. Neurocomputing, 2017, 260: 302 - 312.

[9] Hall M A, Smith L A. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper[C]//Twelfth International Florida Artificial Intelligence Research Society Conference. Orlando, FL, USA: DBLP, 1999: 235 - 239.

[10] Maldonado S, Weber R. A wrapper method for feature selection using support vector machines[J]. Information Sciences, 2009, 179(13): 2208 - 2217.

[11] Archer K J, Kimes R V. Empirical characterization of random forest variable importance measures[J]. Computational Statistics & Data Analysis, 2008, 52(4): 2249 - 2260.

[12] Genuer R, Poggi J M, Tuleau-Malot C, et al. Random forests for big data[J]. Big Data Research, 2017, 9: 28 - 46.

[13] 孙广路, 宋智超, 刘金来, 等. 基于最大信息系数和近似马尔科夫毯的特征选择方法[J]. 自动化学报, 2017, 43(5): 795 - 805.
Sun G L, Song Z C, Liu J L, et al. Feature selection method based on maximum information coefficient and approximate Markov blanket[J]. Acta Automatica Sinica, 2017, 43(5): 795 - 805.

[14] Zhang J, Chen M, Zhao S, et al. Relief F-based EEG sensor selection methods for emotion recognition[J]. Sensors, 2016, 16(10): 1558.

[15] Liu X, Wang L, Zhang J, et al. Global and local structure preservation for feature selection[J]. IEEE Transactions on Neural Networks & Learning Systems, 2017, 25(6): 1083 - 1095.

[16] Zhou Q, Zhou H, Li T. Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features[J]. Knowledge-

Based Systems，2016，95：1 – 11.
[17] Ramrez-Gallego S，Krawczyk B，Woniak M，et al. A survey on data preprocessing for data stream mining：Current status and future directions [J]. Neurocomputing，2017，239(C)：39 – 57.
[18] Guyon I，Elisseeff A. An introduction to variable and feature selection[C]//Joint International Conference on Artificial Neural Networks and Neural Information Processing. Berlin，Germany：Springer-Verlag，2003：737 – 744.
[19] Zhang Y，Ding C，Li T. Gene selection algorithm by combining reliefF and mRMR[J]. BMC Genomics，2008，9(S2)：S27 – S27.
[20] Muriel G，Ioannis D，Sovan L. Review and comparison of methods to study the contribution of variables in artificial neural network models[J]. Ecological Modelling，2003，160(3)：249 – 264.
[21] 王宏威，李国和. 基于属性相似度的连续型特征选择方法[J]. 渤海大学学报(自然科学版)，2014(4)：350 – 355.
Wang H W，Li G H. Continuous feature selection method based on attribute similarity[J]. Journal of Bohai University (Natural Science Edition)，2014(4)：350 – 355.
[22] 邓小龙. 基于距离相关的最小冗余最大相关特征选择法在 QSAR 中的应用[D]. 长沙：湖南农业大学，2016.
Deng X L. Application of minimum redundancy maximum correlation feature selection based on range correlation in QSAR[D]. Changsha：Hunan Agricultural University，2016.
[23] 李捷，陈彦如，杨璐. 基于两阶段组合预测模型的区域物流需求预测[J]. 信息与控制，2018，47(2)：247 – 256.
Li J，Chen Y R，Yang L. Regional logistics demand forecasting based on two-stage combination forecasting model[J]. Information and Control，2018，47(2)：247 – 256.
[24] 姚登举，杨静，詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版)，2014，44(1)：137 – 141.
Yao D J，Yang J，Zhan X J. Feature selection algorithm based on random forest[J]. Journal of Jilin University (Engineering)，2014，44(1)：137 – 141.
[25] Zhou X，Tuck D P. MSVM-RFE：Extensions of SVM-RFE for multiclass gene selection on DNA microarray data[J]. Bioinformatics，2007，23(9)：1106.
[26] Mundra P A，Rajapakse J C. SVM-RFE with MRMR filter for gene selection[J]. IEEE Transactions on Nanobioscience，2010，9(1)：31 – 37.

作者简介

李国和(1965 –)，男，博士，教授，博士生导师. 研究领域为人工智能.
杨绍伟(1993 –)，男，硕士. 研究领域为特征工程.
吴卫江(1971 –)，男，博士生，副教授，硕士生导师. 研究领域为数据挖掘、复杂网络理论及应用.

(上接第 223 页)

[29] Fan M，Ge Z Q，Song Z H. Adaptive Gaussian mixture model-based relevant sample selection for JITL soft sensor development[J]. Industrial & Engineering Chemistry Research，2014，53(51)：19979 – 19986.
[30] Shigemori H，Kano M，Hasebe S. Optimum quality design system for steel products through locally weighted regression model[J]. Journal of Process Control，2011，21(2)：293 – 301.
[31] Kim S，Kano M，Nakagawa H，et al. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection[J]. International journal of pharmaceutics，2011，421(2)：269 – 274.
[32] Hazama K，Kano M. Covariance-based locally weighted partial least squares for high-performance adaptive modeling[J]. Chemometrics and Intelligent Laboratory Systems，2015，146：55 – 62.
[33] Yu J. Online quality prediction of nonlinear and non-Gaussian chemical processes with shifting dynamics using finite mixture model based Gaussian process regression approach[J]. Chemical Engineering Science，2012，82：22 – 30.
[34] Yu J. Multiway Gaussian mixture model based adaptive kernel partial least squares regression method for soft sensor estimation and reliable quality prediction of nonlinear multiphase batch processes[J]. Industrial & Engineering Chemistry Research，2012，51(40)：13227 – 13237.
[35] Figueiredo M A T，Jain A K. Unsupervised learning of finite mixture models[J]. IEEE Transactions on pattern analysis and machine intelligence，2002，24(3)：381 – 396.
[36] Brown G，Wyatt J，Harris R，et al. Diversity creation methods：A survey and categorization[J]. Information Fusion，2005，6(1)：5 – 20.
[37] Zhou Z H. Ensemble methods：Foundations and algorithms[M]. Florida，USA：Chapman and Hall/CRC，2012.
[38] Li G，Aute V，Azarm S. An accumulative error based adaptive design of experiments for offline metamodeling[J]. Structural and Multidisciplinary Optimization，2010，40(1/2/3/4/5/6)：137.

作者简介

潘　贝(1992 –)，女，硕士生. 研究领域为软测量建模与应用.
金怀平(1987 –)，男，博士，讲师，硕士生导师. 研究领域为软测量建模与应用，机器学习与数据挖掘.
杨　彪(1974 –)，男，博士，副教授，硕士生导师. 研究领域为复杂生产过程优化和网络控制，工业智能检测与监控.