基于SAC的特征选择算法

张梦林 李占山
(吉林大学计算机科学与技术学院 长春 130012)
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 特征选择通过移除不相关和冗余的特征来提高学习算法的性能。基于进化算法在求解优化问题时表现出的优越性能,提出FSSAC特征选择方法。新的初始化策略和评估函数使得SAC能将特征选择作为离散空间搜索问题来解决,利用特征子集的准确率指导SAC的采样阶段。在实验阶段,FSSAC结合SVM,J48和KNN分类器,通过UCI数据集完成验证,并与FSFOA,HGAFS,PSO等算法进行了比较。实验结果表明,FSSAC可以提高分类器的分类准确率,且具有良好的泛化性能。除此之外,对FSSAC和其他算法在特征空间维度缩减情况方面做了对比。

关键词 特征选择,SAC,FSSAC,维度缩减
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.011

Feature Selection Algorithm Using SAC Algorithm

ZHANG Meng-lin LI Zhan-shan
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)
(Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun 130012, China)

Abstract Feature selection can improve the performance of learning algorithm with the help of removing the irrelevant and redundant features. As evolutionary algorithm is reported to be suitable for optimization tasks, this paper proposed a new feature selection algorithm FSSAC. The new initialization strategy and evaluation function make FSSAC regard feature selection as a discrete space search problem. The algorithm also uses the accuracy of feature subset to guide the sampling period. In the stage of experiment, FSSAC was combined with the SVM, J48 and KNN, and then it was validated on UCI machine learning datasets by comparing with FSFOA, HGAFS, PSO and so on. The experiments show that FSSAC can improve the classification accuracy of classifier and has good generalization. Besides, FSSAC was also compared with other available methods in dimensionality reduction.

Keywords Feature selection, SAC, FSSAC, Dimensionality reduction

1 引言

当处理数据库中的海量、高维数据时,假设特征数是 n,则评价所有特征子集的时间复杂度是 $O(2^n)$[1],这是难以接受的。特征选择方法是数据挖掘的基础,主要选择在后续学习阶段有使用价值的特征,忽略不相关和冗余的特征。随着高维数据的快速积累(如数字图像、金融时间序列和基因表达微阵列),特征选择在机器学习和数据挖掘领域中成为了重要的预处理阶段。很多文献指出,特征选择提高了机器学习算法的分类器准确率(如KNN分类器[2])。尽管一些学者将不同的搜索策略应用到特征选择过程中,但是大多数算法都存在局部最优化问题且时间开销庞大,因此全局搜索策略在特征选择过程中是十分必要的。

近年出现了进化计算技术,其主要优势在于它的全局搜索策略和可接受的时间开销。文献[3]提出的SAC(Sampling-and-Classification Optimization)是一种较新的进化计算方法,它的计算开销大幅低于其他算法,但是直接用SAC进行特征选择存在一些局限性:1)SAC尚未解决特征选择问题,Qian尝试用SAC解决几个经典的伪布尔问题,且针对不同的问题,SAC有不同的初始化方法,但目前尚未出现针对特征选择问题的初始化策略;2)评估函数的选择存在局限性,原始的SAC评估函数主要用于解决伪布尔问题,不适合用于特征选择。文中提出关于特征选择的SAC初始化策略;考虑到特征选择对分类器的准确率和复杂度有着很大的影响,以分类器在预测数据集上的准确率作为评价特征子集的标准;另外,假设特征空间维度缩减可能会提高分类器的分类准确率或者至少对准确率没有影响。

本文的主要工作有两点:

到稿日期:2017-03-14 返修日期:2017-07-01 本文受吉林省科技发展计划项目(20140101200JC)资助。
张梦林(1991-),男,硕士,主要研究方向为机器学习,E-mail:zhangml5511@163.com;李占山(1966-),男,博士,教授,CCF会员,主要研究方向为约束求解、基于模型的诊断、配置求解与冲突解释,E-mail:zslizsli@163.com(通信作者)。

1)提出一种新的基于 SAC 特征选择算法(FSSAC)和针对特征选择问题特点的 SAC 初始化策略,将特征选择作为一个离散空间搜索问题,以提高 SAC 解决特征选择问题的性能。

2)重新定义评估函数,以分类器在预测数据集上的准确率作为评价特征子集的标准,因此最后选择出的特征子集能够对数据集进行很好的分类。

2 特征选择算法

人们已经对特征选择方法进行了许多研究。依据特征子集的形成方式,特征选择方法可以分为穷举法、启发法和随机法。Almuallim 和 Dietterich 基于穷举法提出的 FOCUS 对整个搜索空间进行搜索,直到找到一个最小的特征子集将训练数据分成单纯的类为止[4-5]。但是 FOCUS 的时间复杂度是指数级的,当特征数非常庞大时,评价所有的特征子集在时间开销上几乎是不可能接受的。因此,当数据集的特征数和实例数都非常庞大时,很少使用穷举法。

特征选择的启发式方法主要包括爬山法、分支界限法、定向搜索和最佳优先算法。爬山算法对局部的特征子集进行评价,找到较好的特征子集,从而能够提高效率。比较典型的爬山算法有 SFS 和 SBS,这两种算法由于不是全局搜索,因此其结果可能不是全局最优的。SFFS(顺序前向浮动搜索)和 SBFS(顺序后向浮动搜索)解决了这一问题[6],其允许在局部搜索过程中回溯刚才遍历的特征。

随机方法是近几年较新的特征选择方法,可以细分为完全随机方法和概率随机方法。完全随机方法是指特征子集的产生不依赖任何概率,而概率随机方法按照用户给定的概率产生某个特征子集。虽然随机方法的搜索空间是 $O(2^n)$,但是可以通过设置最大的迭代次数来限制其搜索空间小于 $O(2^n)$。近几年许多学者将遗传算法(GA)、粒子群算法(PSO)和蚁群算法(ACO)等进化计算算法应用到特征选择中,在时间效率方面都表现出令人非常满意的结果。如 Zhu 等提出将遗传算法和局部搜索方法结合在一起的特征选择算法[7]。Ghaemi 用解决连续问题的森林优化算法离散化求解特征选择问题[8],但是随机方法存在较高的不确定性,只有当循环次数较大时才可能找到比较好的结果。在随机搜索策略中,可能需要对一些参数进行设置,参数选择的合适与否将直接影响最终实验结果的好坏,因此参数选择也是一个比较大的问题。

上述方法评价算法的标准是在计算时间上可接受,或者对特征选择的优化较好。为了满足实际应用的需要,针对特征选择问题的研究需要进一步提出性能更加优越的算法。为此,文中提出了 FSSAC 算法,其通过提高分类准确率进一步优化特征选择的性能,搜索空间远小于 $O(2^n)$,算法的执行时间也是可接受的;除此之外,对多个数据集的测试也表明 FSSAC 具有很好的泛化性能。

3 FSSAC 方法

3.1 Sampling-and-Classification(SAC)优化算法

近些年来对进化算法的各种实现几乎都是基于一个通用的结构,即包含采样和模型建立的循环,这种通用的结构被称为 Sampling-and-Learning(SAL)[9]框架,如算法 1(求解最小值问题)所示。

算法 1 The SAL framework

输入:$\alpha^* > 0$ 表示近似水平;$\lambda \in (0,1)$ 表示平衡参数;$T \in N^+$ 表示迭代次数;$m_0, \cdots, m_T \in N^+$ 表示采样规模;L 表示学习算法;T 表示假设的分布变换

输出:$x^{\sim}$

步骤:

1. 从分布 X 中随机采样,构成集合 $S_0 \in \{x_1, \cdots, x_{m_0}\}$;
2. 令 $x^{\sim} = \arg\min_{x \in S_0} f(x)$;
3. 初始化假设 h_0;
4. $T_0 = \emptyset$;
5. For t=1 to T do
6. 构造 $T_t = \{(x_1, y_1), \cdots, (x_{m_{t-1}}, y_{m_{t-1}})\}$,其中 $x_i \in S_{t-1}, y_i = f(x_i)$;
7. 在学习阶段,$h_t = L(T_t, T_{t-1}, h_{t-1}, t)$;
8. 根据 T_t 初始化 S_t;
9. For i=1 to m_t do
10. Sample x_i from $\begin{cases} T_{ht}, & \text{概率 } \lambda \\ U_X, & \text{概率 } 1-\lambda \end{cases}$;
11. $S_t = S_t \cup \{x^{\sim}\}$;
12. End for
13. $x^{\sim} = \arg\min_{x \in St \cup \{x^{\sim}\}} f(x)$;
14. End for
15. Return $x^{\sim}$.

进化计算通常在产生解的过程中利用一些启发式,这样可以降低时间复杂度和空间复杂度[10],这个思想被应用在 SAL 的采样阶段。同样,它们也会根据产生解的质量来指导下一次的采样,这个过程被应用到 SAL 的学习阶段。SAL 框架的一种简单实现方式是在它的学习阶段使用一个分类算法,并将其称为 Sampling-and-classification(SAC)[3]算法。SAC 包括 3 个阶段:1)全局初始化采样;2)学习阶段;3)全局采样。

在全局初始化采样阶段,SAC 随机地从整体样本集 X 中均匀同分布采样构成初始集合 S_0,并计算 S_0 中的目标函数 $f(x)$的最优值。下一个阶段是学习阶段,在该阶段中,SAC 根据上次迭代中采样的数据集构造训练数据集,并用该训练数据集训练学习器。将训练好的学习器用来指导下一阶段的采样。最后一个阶段是为下一次迭代进行采样,它是以概率 λ 在 T_{ht} 分布中采样,以概率 $1-\lambda$ 在全局 U_x 中采样,其中 T_{ht} 是训练后的分类器对全局样本进行分类所获得的正类样本分布情况。

文中尝试将特征选择问题转化为离散空间搜索问题并用 SAC 来解决,具体的伪代码见算法 2。根据特征选择问题的特点,提出新的 SAC 初始化策略。考虑到特征选择对分类器的准确率和复杂度有着很大的影响,将在预测数据集上的准确率作为评价特征子集的标准。

算法 2 FSSAC

输入:f 表示特征子集的准确率;$\alpha_1 > \cdots > \alpha_T$ 表示用来标记的阈值;$\lambda \in (0,1)$ 表示平衡参数;$T \in N^+$ 表示迭代次数;$m_t \in N^+$ 表示采

样数量；H 表示分类器集合；C 表示二元分类算法

输出：$x^{\sim}$

步骤：

1. 从分布 X 中均匀采样，构成集合 $S_0 \in \{x_1, \cdots, x_{m_0}\}$，其中分布 X 是所有特征子集的整体分布情况，采样范围是 $(1-2^{n-1})$，其中每个整数代表一个特征子集；
2. 令 $x^{\sim}=\arg\max_{x \in S_0} f(x)$；
3. for t=1 to T do
4. 构造 $B_t=\{(x_1, y_1), \cdots, (x_{m_{t-1}}, y_{m_{t-1}})\}$，其中 $x_i \in S_{t-1}$，$y_i=\mathrm{sign}[f(x_i)-\alpha_t]$；
5. 分类器训练 $h_t=C(B_t)$，其中 $h_t \in H$；
6. $S_t=\varnothing$；
7. for i=1 to m_t do
8. Sample x_i from $\begin{cases} U_{Dht}, & \text{概率 } \lambda \\ U_X, & \text{概率 } 1-\lambda \end{cases}$；
9. $S_t=S_t \cup \{x^{\sim}\}$；
10. End for
11. $x^{\sim}=\arg\max_{x \in S_t \cup \{x^{\sim}\}} f(x)$；
12. End for
13. Return $x^{\sim}$.

3.2 FSSAC

本节主要从初始化采样策略、学习阶段、采样阶段 3 个方面详细介绍 FSSAC 特征选择算法。

3.2.1 初始化采样策略

FSSAC 的开始阶段是初始化采样，首先从整个样本数据集中均匀采样。用一个二进制串来表示特征的选取情况。假设整个样本集共有 n 个特征，则该二进制串的大小是 n，每一个二进制用 0 或者 1 来表示。若该二进制位是 1，则该位对应的特征被选取并用在后面的学习器学习步骤中；若是 0，则该位对应的特征在后面的学习器学习阶段并未被选取。为方便取样，将每一个二进制串转换成一个整数。当样本数据集的特征总数为 n 时，可以选择的特征子集数量为 $[1-(2^n-1)]$，因此在 FSSAC 的初始化阶段其实是从 $1-(2^n-1)$ 中均匀取 m 个整数，每个整数代表一个特征子集。

3.2.2 学习阶段

FSSAC 的学习阶段是在一个循环中完成的。在每一次迭代中，使用一个目标函数去估计当前数据集中的每个特征子集，这里分类器对于每个特征子集的分类准确率被选择作为目标函数；然后构造一个二元的分类数据集 B_t，根据 $\mathrm{sign}[f(x)-\alpha_t]$ 将特征子集标记为正类或者负类，其中 $f(x)$ 代表分类器对于特征子集 x 的分类准确率，α_t 是第 t 次迭代中用户预设的准确率阈值。学习阶段的主要目的是用数据集 B_t 训练一个二元分类器，然后用二元分类器近似一个区域 $D_{\alpha t}=\{x \in X \mid f(x)>\alpha_t\}$。

3.2.3 采样阶段

在 FSSAC 的采样过程中，由特征子集构成的集合是通过 U_{Dht}（分类器对所有特征子集进行分类，由正类构成的均匀分布）以概率 λ 采样和以概率 $1-\lambda$ 在集合 U_X（在整个特征子集上的均匀分布）采样得到的。值得强调的是，在分布 U_{Dh} 中采样可能暗示着从一个质量比较好的区域中采样，而这个区域是由二元分类器 h 学习得到的。参数 λ 被用来平衡局部搜索和全局搜索。来自任意区域的采样对最后的实验结果具有直接的影响，在 FSSAC 特征选择方法中，由于在训练数据集 B_t 的构造过程中正类的数据通过 $\mathrm{sign}[f(x)-\alpha_t]$ 来进行标记，只有大于 α_t 的特征子集才被标记为正类，随着迭代次数的增加，α_t 也逐渐变大，相应的被标记为正类的特征子集代表的准确率也不断增大。由于在采样阶段以较大的概率在 D_h 分布上采样，因此在正类区域 D_h 上采样是最直接有效的。

4 实验与分析

实验阶段使用的是公开的 scikit-learn 包，它是基于 python 语言的机器学习工具。所有实验均在 DELL 机器上执行，其配置为 Intel Core i7 CPU(3.60 GHz)，8 GB 内存；主要的编程语言是 python。

4.1 数据集

将提出的 FSSAC 算法在 9 个数据集上进行验证，这些数据集包含 5 个小维数据集、2 个中维数据集和 2 个高维数据集，如表 1 所列。其中，Ionosphere，Glass，Cleveland，Vehicle，Dermatology，Sonar，Wine，Segmentation 来自 UCI 数据集，SRBCT 来自微阵列数据集。对于每个数据集，根据对比算法将其按照 10-折交叉验证、2-折交叉验证和 70%用于训练而 30%用于预测集的方式进行划分。将测试集的分类准确率(CA)和维度缩减(DR)作为评价 FSSAC 及对比算法的指标。分类准确率的具体定义如式(1)所示，其中 N_CC 是正确分类的实例数，N_AS 是数据集的整体实例数。维数缩减率的计算公式如式(2)所示，其中 N_SF 是选择特征的数量，N_AF 是数据集上所有的特征数量。

$$CA=N\_CC/N\_AS \tag{1}$$

$$DR=1-(N\_SF/N\_AF) \tag{2}$$

FSSAC 算法主要涉及到的平衡参数 λ 和迭代次数 T 在实验中被设置为 $T=6$，$\lambda=0.7$。

表 1　5 个小维数据集、2 个中维数据集和 2 个高维数据集

Table 1　Five small dimensional data sets，two middle dimensional data sets and two high-dimensional data sets

Dataset	# Feature	# Instances	# Class
Vehicle	18	846	4
Cleveland	13	303	5
Dermatology	34	366	6
Ionosphere	34	351	2
Glass	9	214	7
Wine	13	178	3
Segmentation	19	2310	7
SRBCT	2308	63	4
Sonar	60	208	2

4.2 实验结果对比

将提出的 FSSAC 算法与其他经典算法进行比较。用于对比的特征选择算法有：Ghaemi 利用分类器准确率作为评价标准提出的 FSFOA[8]；Hu 等基于邻居的软间隔评价特征子集而提出的 NSM[11]；Moustakidis 等提出的 SVM-FuzCoc[12]，其使用了模糊的互补准则的 SVM；Huang 提出的 FS 混合基因算法(HGAFS)[13]，其和 SFS，SBS，SFFS[12] 利用了

互信息评价特征子集；Zhu 等利用邻居有效信息率评价准则提出的 FS-NEIR[7]；Tabakhi 等基于特征之间的相似性评价准则提出的 UFSACO[14]和 Xue 等提出的将分类器准确率和特征子集规模作为评价准则的 PSO(4-2)[15]。这些算法的具体信息如表 2 所列。

表 2 对比算法的详细信息

Table 2 Details of the comparison algorithm

算法名称	数据集划分	描述/发表年份
FSFOA	70-30 10-fold 2-fold	基于森林优化算法的特征选择[8]/2016
SFS,SBS,SFFS	70-30	贪婪的爬山算法[12]/2010
NSM	10-fold	相邻的软间隔[11]/2010
SVM-FuzCoc	70-30	基于 SVM 的特征选择[12]/2010
HGAFS	2-fold	混合遗传算法[13]/2007
FS-NEIR	10-fold	相邻有效信息比[7]/2013
UFSACO	70-30	基于 ACO 的无监督的特征选择算法[14]/2014
PSO(4-2)	10-fold	基于粒子群优化的特征选择算法[15]/2013

FSSAC 和表 2 所列对比算法的分类准确率、维度缩减情况如表 3 所列。在每个数据集中，不同分类器对应的最优的分类准确率和维度缩减用粗体标出。实验中使用的分类器主要包括 KNN(1-NN,3-NN,5-NN),C4.5 和 Rbf-SVM。在学习阶段选择的二元分类器为 1-NN,rbf-svm 和 gini。

在表 3 中，数据集 Glass 和 Segmentation 在相同的划分方法和分类器的情况下，FSSAC 算法的准确率是最高的。所选择的对比算法都是基于遗传算法、蚁群优化算法和森林优化算法提出的。FSSAC 在局部搜索和全局搜索特征空间中的最优特征子集的性能比其他基于进化计算的特征选择算法更优。这表明 FSSAC 通过移除这两个数据集上的冗余特征提高了 KNN，J48 和 SVM 分类器的准确率。在数据集 Wine，Ionosphere，Dermatology 和 Sonar 中，FSSAC 可以表现出比大多数对比算法更好的分类性能；在这些数据集中，除不及 1 个对比算法外，FSSAC 都表现出了较优的分类能力。在 Sonar 数据集上，除了分类器 SVM 的准确率比 HGAFS 低以外，在相同的划分方法和分类器的条件下 FSSAC 的准确率比对比算法高出 15%，甚至在数据集 70%-30%的划分条件下 FSSAC 比 SBS 高出 30%。在 Cleveland 数据集上，Rbf-SVM-FuzCoc 的分类性能稍好，FSSAC 和其他对比算法的分类准确率在 52%左右。同样，在 Vehicle 数据集上，FSSAC 在分类器 J48 上的准确率要比 FSFOA 和 FS-NEIR 高。当分类器是 5-NN 和 SVM 时，PSO(4-2)和 HGAFS 分别是对应的分类器上最优的算法，FSSAC 和其他对比算法的分类准确率在 73%和 55%左右。FSSAC 在数据集 SRBCT 上的分类性能比 FSFOA 和 Rbf-SVM-FuzCoc 稍差，但是 FSSAC 的特征空间维度缩减要比其他两个算法优越。通过 SRBCT 数据集的特点可以看出，SRBCT 有 2308 个特征，实例数只有 63 个，即特征数远大于实例数，这会致使在庞大的特征空间中选择合适的特征变得很困难，并且由于实例数很少，分类器很难选择出具有相关性的特征，从而使分类准确率降低；同时，该数据集的划分方法是 70%-30%，导致实验结果更坏，这是因为在训练阶段数据集在一定程度上被忽略了。

表 3 FSSAC 与其他算法的分类准确率和维度缩减对比

Table 3 Classification accuracy and dimension reduction of FSSAC compared with other algorithms

Cleveland	Accuracy/%	DR/%	Classifier
FSSAC	52.87(70%-30%)	69.23	gini/1-NN
	49.43(70%-30%)	61.54	1-NN/1-NN
	48.28(70%-30%)	61.54	Rbf-svm/1-NN
FSFOA	55.55(70%-30%)	**71.42**	1-NN
RBF-SVM-FuzCoc	**61.01(70%-30%)**	46.1	1-NN
SFS	51.79(70%-30%)	47.7	1-NN
SBS	54.8(70%-30%)	38.5	1-NN
SFFS	49.55(70%-30%)	53.8	1-NN
Segmentation	**Accuracy/%**	**DR/%**	**Classifier**
FSSAC	96.02(10-fold)	37.89	Rbf-svm/3-NN
	96.80(10-fold)	39.47	1-NN/3-NN
	96.89(10-fold)	34.74	gini/3-NN
FSFOA	96.2(10-fold)	30.00	3-NN
NSM	95(10-fold)	**63.15**	3-NN
Vehicle	**Accuracy/%**	**DR/%**	**Classifier**
FSSAC	**76.23(10-fold)**	47.22	1-NN/J48
	76.23(10-fold)	49.44	Rbf-svm/J48
	76.95(10-fold)	42.78	gini/J48
FSFOA	73.04(10-fold)	31.57	J48
FS-NEIR	70.98(10-fold)	**50.00**	J48
FSSAC	72.42(70%-30%)	47.21	gini/5-NN
	73.02(70%-30%)	52.78	1-NN/5-NN
	72.22(70%-30%)	44.44	Rbf-svm/5-NN
FSFOA	73.98(70%-30%)	50	5-NN
PSO(4-2)	**85.3(70%-30%)**	**68.4**	5-NN
FSSAC	50.24(2-fold)	63.89	Rbf-svm/ svm
	55.44(2-fold)	**66.67**	1-NN/Rbf-svm
	50(2-fold)	61.11	gini/Rbf-svm
FSFOA	62.41(2-fold)	47.22	Rbf-svm
HGAFS	**76.36(2-fold)**	38.89	Rbf-svm
Dermatology	**Accuracy/%**	**DR/%**	**Classifier**
FSSAC	**98.90(10-fold)**	44.12	Rbf-svm/J48
	98.90(10-fold)	47.35	gini/J48
	99.17(10-fold)	46.18	1-NN/J48
FSFOA	96.99(10-fold)	21.42	J48
FS-NEIR	93.95(10-fold)	**70.58**	J48
FSSAC	95.37(70%-30%)	54.55	1-NN /1-NN
	96.30(70%-30%)	32.35	Rbf-svm/1-NN
	96.30(70%-30%)	51.52	gini/1-NN
FSFOA	**97.27(70%-30%)**	45.71	1-NN
SFS	94.02(70%-30%)	44.7	1-NN
SBS	91.78(70%-30%)	58.23	1-NN
SFFS	93.7(70%-30%)	62.35	1-NN
RBF-SVM-FuzCoc	94.11(70%-30%)	**64.7**	1-NN
FSSAC	**96.30(70%-30%)**	39.40	gini/J48
	96.30(70%-30%)	42.42	1-NN/J48
	96.30(70%-30%)	**47.06**	Rbfsvm/J48
FSFOA	90.09(70%-30%)	44.11	J48
UFSACO	95.28(70%-30%)	26.47	J48
Sonar	**Accuracy/%**	**DR/%**	**Classifier**
FSSAC	62.02(2-fold)	55.00	1-NN/svm
	61.54(2-fold)	60.00	gini/Rbf-svm
	60.10(2-fold)	55.83	Rbfsvm/svm
FSFOA	65.86(2-fold)	54.09	Rbf-svm
HGAFS	**87.02(2-fold)**	**75.00**	Rbf-svm
FSSAC	**89.40(10-fold)**	48.00	1-NN/J48
	89.88(10-fold)	51.83	gini/J48
	90.36(10-fold)	51.00	Rbf-svm/J48
FSFOA	82.69(10-fold)	52.45	J48
FS-NEIR	75.97(10-fold)	**91.66**	J48
FSSAC	**95.00(70%-30%)**	55.00	svm/1-NN
	95.00(70%-30%)	45.00	1-NN/1-NN
	95.00(70%-30%)	48.33	gini/1-NN
FSFOA	85.43(70%-30%)	57.37	1-NN
RBF-SVM-FuzCoc	73.17(70%-30%)	**68.33**	1-NN
SFS	66.43(70%-30%)	61.33	1-NN
SBS	62.2(70%-30%)	45.33	1-NN
SFFS	64.55(70%-30%)	61.33	1-NN
FSSAC	**95.00(70%-30%)**	50.00	gini/5-NN
	95.00(70%-30%)	43.33	svm/5-NN
	90.00(70%-30%)	50.00	1-NN/5-NN
FSFOA	86.98(70%-30%)	44.26	5-NN
PSO(4-2)	78.16(70%-30%)	**81.26**	5-NN

（续表）

Ionosphere	Accuracy/%	DR/%	Classifier
	96.00(10-fold)	46.47	Rbf-svm/J48
FSSAC	**96.86(10-fold)**	44.71	1-NN/J48
	96.29(10-fold)	51.12	gini/J48
FSFOA	93.16(10-fold)	68.57	J48
FS-HEIR	92.59(10-fold)	**82.35**	J48
	93.15(10-fold)	56.76	gini/3-NN
FSSAC	**92.58(10-fold)**	53.82	1-NN/3-NN
	92.29(10-fold)	52.65	Rbf-svm/3-NN
FSFOA	92.3(10-fold)	61.76	3-NN
NSM	92(10-fold)	**88.23**	3-NN
	90.59(10-fold)	56.76	Rbf-svm/5-NN
FSSAC	**90.87(10-fold)**	51.76	1-NN/5-NN
	90.58(10-fold)	51.47	gini/5-NN
FSFOA	89.43(10-fold)	54.28	5-NN
PSO(4-2)	87.27(10-fold)	**90.41**	5-NN
	92.02(2-fold)	50.00	gini/Rbf-svm
FSSAC	92.31(2-fold)	60.29	1-NN/Rbf-svm
	92.02(2-fold)	41.18	svm/Rbf-svm
FSFOA	**94.58(2-fold)**	57.14	RBF/SVM
HGAFS	92.76(2-fold)	**82.35**	Rbf-Rbf-svm
	95.29(70%-30%)	**58.82**	gini/J48
FSSAC	**95.29(70%-30%)**	44.12	1-NN/J48
	95.33(70%-30%)	52.94	Rbf-svm/J48
FSFOA	95.12(70%-30%)	47.05	J48
UFSACO	88.61(70%-30%)	11.17	J48
	92.38(70%-30%)	52.94	Rbf-svm/1-NN
FSSAC	**90.48(70%-30%)**	44.12	1-NN/1-NN
	90.48(70%-30%)	50.00	gini/1-NN
FSFOA	89.52(70%-30%)	54.28	1-NN
RBF-SVM-FuzCoc	89.46(70%-30%)	**88.23**	1-NN
SFS	87.75(50%-50%)	65.88	1-NN
SBS	84.61(50%-50%)	77.64	1-NN
SFFS	88.32(50%-50%)	75.29	1-NN

Glass	Accuracy/%	DR/%	Classifier
	79.85(10-fold)	44.44	1-NN/J48
FSSAC	**79.70(10-fold)**	42.22	Rbf-svm/J48
	77.27(10-fold)	39.44	gini/J48
FSFOA	75.7(10-fold)	**50.00**	J48
FS-NEIR	68.53(10-fold)	22.22	J48
	74.60(70%-30%)	22.22	J48/1-NN
FSSAC	**74.60(70%-30%)**	11.11	Rbf-svm/1-NN
	73.02(70%-30%)	22.22	1-NN /1-NN
FSFOA	71.88(70%-30%)	**40.00**	1-NN
RBF-SVM-FuzCoc	73.36(70%-30%)	33.33	1-NN
SFS	72.24(70%-30%)	26.66	1-NN
SBS	71.77(70%-30%)	37.77	1-NN
SFFS	71.77(70%-30%)	37.77	1-NN
	67.29(2-fold)	38.89	1NN/Rbfsvm
FSSAC	67.76(2-fold)	38.89	svm/Rbfsvm
	68.69(2-fold)	38.89	gini/Rbf-svm
FSFOA	68.22(2-fold)	**60.00**	Rbf-svm
HGAFS	65.51(2-fold)	44.44	Rbf-svm

SRBCT	Accuracy/%	DR/%	Classifier
	94.44(70%-30%)	**99.87**	Rbf-svm/1-NN
FSSAC	94.44(70%-30%)	**99.87**	1-NN/1-NN
	94.44(70%-30%)	**99.87**	gini/1-NN
FSFOA	94.73(70%-30%)	49.06	1-NN
RBF-SVM-FuzCoc	**98.88(70%-30%)**	98.57	**1-NN**

（续表）

Wine	Accuracy/%	DR/%	Classifier
	99.44(10-fold)	47.69	gini/J48
FSSAC	**99.44(10-fold)**	47.69	1-NN/J48
	97.78(10-fold)	47.69	Rbf-svm/J48
FSFOA	96.06(10-fold)	21.42	J48
FS-NEIR	95.04(10-fold)	**61.53**	J48
	99.44(10-fold)	45.38	gini/3-NN
FSSAC	97.78(10-fold)	53.85	1-NN/3-NN
	96.67(10-fold)	50.77	Rbf-svm/3-NN
FSFOA	98.87(10-fold)	42.58	3-NN
NSM	98.00(10-fold)	**53.84**	3-NN
	98.08(70%-30%)	53.85	Rbfsvm/1NN
FSSAC	**99.99(70%-30%)**	33.33	1-NN/1-NN
	98.08(70%-30%)	**66.67**	gini/1-NN
FSFOA	98.07(70%-30%)	50.00	1-NN
RBF-SVM-FuzCoc	97.12(70%-30%)	53.84	1-NN
SFS	97.69(70%-30%)	35.38	1-NN
SBS	94.77(70%-30%)	46.15	1-NN
SFFS	96.56(70%-30%)	36.92	1-NN
	97.80(70%-30%)	33.33	gini/J48
FSSAC	**97.53(70%-30%)**	33.33	1-NN/J48
	96.70(70%-30%)	41.67	Rbf-svm/J48
FSFOA	96(70%-30%)	57.14	J48
UFSACO	95.08(70%-30%)	**61.53**	J48
	98.08(70%-30%)	**69.23**	Rbf-svm/5-NN
FSSAC	**99.99(70%-30%)**	66.67	1-NN/5-NN
	99.99(70%-30%)	41.67	gini/5-NN
FSFOA	99.2(70%-30%)	30.76	5-NN
PSO(4-2)	95.26(10-fold)	51.6	5-NN
	94.38(2-fold)	**58.33**	gini/Rbf-svm
FSSAC	94.38(2-fold)	50.00	1-NN/Rbf-svm
	94.38(2-fold)	57.69	Rbf-svm/ svm
FSFOA	96.06(2-fold)	37.17	Rbf-svm
HGAFS	**98.31(2-fold)**	53.85	Rbf-svm

通过比较表 3 的维度缩减发现，FSSAC 在某些数据集上并没有表现出非常大的优势，正如前面所述，FSSAC 算法使用分类器的准确率作为评价函数进行训练数据的构造，而没有考虑到在特征选择过程中特征空间维度缩减的情况。另外，表 3 也说明了与经典的基于进化计算的特征选择算法相比，FSSAC 具有非常好的泛化性能，将特征选择作为一个优化问题可以达到预期的效果。

结束语　本文提出了 FSSAC 特征选择算法，该方法通过提出新的初始化策略和评估函数，使 SAC 能够解决特征选择问题。为了研究 FSSAC 算法的性能，选择了 UCI 中几个比较著名的数据集和较有代表性的对比算法进行实验。所选择的对比算法包括遗传算法、蚁群算法、粒子群优化算法和森林优化算法。通过分析 9 个数据集的实验结果发现，FSSAC 普遍提高了数据集的分类准确率，并且具有很好的泛化性能。实验结果表明，对于解决特征选择问题，FSSAC 是一种有效的特征搜索方法。

接下来将尝试在采样阶段不用特征子集的分类准确率作为唯一的评价标准，结合特征子集的区分度来指导下一迭代步骤中的采样阶段；在面对特征数庞大而实例数较少的数据集时，FSSAC 的性能值得进一步研究，因为在现阶段，不管是机器学习还是数据挖掘，庞大的数据集（具有非常大的特征空间和实例）仍然是难以解决的难题；另外，对于 FSSAC 算法中的维度缩减，可以通过增加多目标评估函数来实现这一目的，使评估函数可以同时对分类准确率和维度缩减情况进行优化。

参 考 文 献

[1] TAN K C,TEOH E J,YU Q,et al. A hybrid evolutionary algorithm for attribute selection in data mining[J]. Expert Systems with Applications,2009,36(4):8616-8630.

[2] HALL M A. Correlation-Based Feature Selection for Machine Learning[D]. Hamilton:The University of Waikato,1999.

[3] HONG Q,YANG Y. On Sampling-and-Classification Optimization in Discrete Domains[C]// IEEE Congress on Evolutionary Computation. IEEE,2016.

[4] ALMUALLIM H,DIETTERICH T G. Learning Boolean concepts in the presence of many irrelevant features[J]. Artificial Intelligence,1994,69(1-2):279-305.

[5] ALMUALLIM H,DIETTERICH T G. Learning with many irrelevant features[C]// National Conference on Artificial Intelligence. AAAI Press,1991:547-552.

[6] PUDIL P,NOVOVI,KITTLER J. Floating search methods in feature selection[J]. Pattern Recognition Letters,1994,15(11):1119-1125.

[7] ZHU W,SI G,ZHANG Y,et al. Neighborhood effective information ratio for hybrid feature subset evaluation and selection [J]. Neurocomputing,2013,99:25-37.

[8] GHAEMI M,FEIZI-DERAKHSHI M R. Feature selection using Forest Optimization Algorithm[J]. Pattern Recognition,2016,60:121-129.

[9] YU Y,QIAN H. The sampling-and-learning framework:A statistical view of evolutionary algorithms[C]// Evolutionary Computation. IEEE,2014:149-158.

[10] SUTTON A M,NEUMANN F. A Parameterized Runtime Analysis of Evolutionary Algorithms for the Euclidean Traveling Salesperson Problem[C]// AAAI Conference on Artificial Intelligence. 2012:595-628.

[11] HU Q,CHE X,ZHANG L,et al. Feature evaluation and selection based on neighborhood soft margin[J]. Neurocomputing,2010,73(10-12):2114-2124.

[12] MOUSTAKIDIS S P,THEOCHARIS J B. SVM-FuzCoC:A novel SVM-based feature selection method using a fuzzy complementary criterion[J]. Pattern Recognition,2010,43(11):3712-3729.

[13] HUANG J,RONG P. A Hybrid Genetic Algorithm for Feature Selection Based on Mutual Information[J]. Pattern Recognit. Lett.,2007,28(13):1825-1844.

[14] TABAKHI S,MORADI P,AKHLAGHIAN F. An unsupervised feature selection algorithm based on ant colony optimization [J]. Engineering Applications of Artificial Intelligence,2014,32(6):112-123.

[15] XUE B,ZHANG M,BROWNE W N. Particle swarm optimisation for feature selection in classification:Novel initialisation and updating mechanisms [J]. Applied Soft Computing,2014,18(C):261-276.

(上接第 52 页)

[7] SHI W S,SUN H,CAO J,et al. Edge Computing-An Emerging Computing Model for Internet of Everything Era[J]. Journal of Computer Research and Development,2017,54(5):907-924. (in Chinese)
施巍松,孙辉,曹杰,等. 边缘计算:万物互联时代新型计算模型[J]. 计算机研究与发展,2017,54(5):907-924.

[8] XU R,GUO J,DENG L. A database security gateway to the detection of SQL attacks[C]// International Conference on Advanced Computer Theory and Engineering. IEEE, 2010: 537-540.

[9] MURRAY A T,MATISZIW T C,GRUBESIC T H. A Methodological Overview of Network Vulnerability Analysis[J]. Growth & Change,2008,39(4):573-592.

[10] CHADWICK D W,BASDEN A. Evaluating Trust in a Public Key Certification Authority[J]. Computers & Security,2001,20(7):592-611.

[11] BRICKELL E,CAMENISCH J,CHEN L. Direct anonymous attestation[C]// ACM Conference on Computer and Communications Security. ACM,2004:132-145.

[12] ZOHAR A. Bitcoin[J]. Communications of the Acm,2015,58(9):104-113.

[13] Ethereum[EB/OL]. https://www.ethereum.org.

[14] LI W,SFORZIN A,FEDOROV S,et al. Towards Scalable and Private Industrial Blockchains[C]// ACM Workshop on Blockchain,Cryptocurrencies and Contracts. ACM,2017:9-14.

[15] UNDERWOOD S. Blockchain Beyond Bitcoin[J]. Communications of the Acm,2016,59(11):15-17.

[16] HE P,YU G,ZHANG Y F,et al. Surveyon Blockchain Technology and Its Application Prospect[J]. Computer Science,2017,44(4):1-7. (in Chinese)
何蒲,于戈,张岩峰,等. 区块链技术与应用前瞻综述[J]. 计算机科学,2017,44(4):1-7.

[17] DPoS[EB/OL]. http://8btc.com/article-3759-1.html.

[18] Bitshares[EB/OL]. http://www.btsabc.org.

[19] DOLEV D,YAO A. On the Security of Public Key Protocols [J]. IEEE Transactions on Information Theory,1983,29(2):198-208.

[20] ZHU L H,GAO F,SHEN M,et al. Survey of block chain privacy protection[J]. Journal of Computer Research and Development,2017,54(10):2170-2186. (in Chinese)
祝烈煌,高峰,沈蒙,等. 区块链隐私保护研究综述[J]. 计算机研究与发展,2017,54(10):2170-2186.

[21] LIN C,PENG X H. Research on trusted network[J]. Chinese Journal of Computers,2005,28(5):751-758. (in Chinese)
林闯,彭雪海. 可信网络研究[J]. 计算机学报,2005,28(5):751-758.