基于文化基因算法和最小二乘支持向量机的安全数据特征处理方法

马媛媛[1] 施永益[2] 张 宏[3] 林 棋[3] 李千目[3]
(全球能源互联网研究院 南京 210003)[1] (国网浙江省电力公司 杭州 310013)[2]
(南京理工大学计算机科学与工程学院 南京 210094)[3]

摘 要 随着各类生物智能演化算法的日益成熟,基于演化技术及其混合算法的特征选择方法不断涌现。针对高维小样本安全数据的特征选择问题,将文化基因算法(Memetic Algorithm,MA)与最小二乘支持向量机(Least Squares Support Vector Machine,LS-SVM)进行结合,设计了一种封装式(Wrapper)特征选择方法(MA-LSSVM)。该方法利用最小二乘支持向量机易于求解的特点来构造分类器,以分类的准确率作为文化基因算法寻优过程中适应度函数的主要成分。实验表明,MA-LSSVM可以较高效地、稳定地获取对分类贡献较大的特征,降低了数据维度,提高了分类效率。

关键词 特征选择,文化基因算法,最小二乘支持向量机,稳定性
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.03.049

Feature Processing Approach Based on MA-LSSVM in Safety Data

MA Yuan-yuan[1] SHI Yong-yi[2] ZHANG Hong[3] LIN Qi[3] LI Qian-mu[3]
(State Grid Smart Grid Research Institute, Nanjing 210003, China)[1]
(Zhejiang Electric Power Corporation, Hangzhou 310013, China)[2]
(School of Computer Science and Engineering, Nanjing University of Science & Technology, Nanjing 210094, China)[3]

Abstract With all kinds of biological intelligent evolutionary algorithms become increasingly mature, feature selection methods based on the evolutionary technology and its hybrid algorithm are emerging. According to the feature selection problem of the high dimensional small sample safety data, this paper combined memetic algorithm (MA) and least squares support vector machines (LS-SVM) to design a kind of wrapper feature selection method (MA-LSSVM). The proposed method utilizes the specialty of least squares support vector machine being easy to search optimal solution to construct classifier, then regards classification accuracy as the main component of memetic algorithm fitness function in the optimization process. The experimental results demonstrate that MA-LSSVM can be more efficient and stable to obtain features larger contribution to the classification precision, and can reduce the data dimension and improve the classification efficiency.

Keywords Feature selection, Memetic algorithm, Least squares support vector machine, Stability

1 介绍

随着计算机技术和网络的不断发展以及各种信息的急剧膨胀,如何有效地从庞大的数据中寻找出有效的信息已经成为人们不可逃避的现实问题。在信息安全方面,膨胀的信息带来各式各样的高维小样本数据。高维小样本问题有以下两种情况:1)各类训练样本的数量远小于特征空间的维度;2)训练样本数量尽管大于特征空间的维度,但两者属于同一数量级[1]。大多数高维小样本数据的特征空间存在大量的冗余特征和噪声特征,这些特征一方面可能降低分类器对数据预测的精度,另一方面也会大幅增加训练的时间和空间复杂度,导致众所周知的"维度灾难"。为了从大量的高维小样本数据中高效地挖掘出有用的知识,特征选择已成为信息安全领域高维小样本数据分析处理中的关键性问题。

目前,特征选择方法呈现出多样化和综合性的研究趋势。根据特征选择方法是否独立于后续的分类算法,将特征选择方法的类型分为过滤式(Filter)、封装式(Wrapper)和嵌入式(Embedded)3种[2]。过滤式方法与后续的分类算法无关,直接用训练样本数据在统计方面的性能来进行特征的评估和选择;封装式方法利用分类器的分类性能作为特征子集评价优

到稿日期:2015-11-06 返修日期:2016-04-01 本文受国家电网公司2015年科技项目(SGRIXTKJ[2015]216)资助。
马媛媛(1978—),女,高级工程师,主要研究领域为信息安全;**施永益**(1969—),男,主要研究领域为企业信息化、信息安全研究及应用;**张 宏**(1956—),男,博士,教授,主要研究领域为网络故障诊断与数据挖掘、信息安全理论与技术,E-mail:zhhong@mail.njust.edu.cn;**林 棋**(1992—),男,硕士,主要研究领域为网络安全、数据挖掘,E-mail:380359384@qq.com(通信作者);**李千目**(1979—),男,博士,教授,主要研究领域为信息安全、大数据处理系统,E-mail:qianmu@mail.njust.edu.cn。

劣的标准或者标准的一部分；嵌入式方法将特征选择算法自身作为组成部分嵌入到分类算法中。过滤式特征选择算法独立于后续的机器学习算法，计算代价相对较小，所以运行速度快，但是特征选择的效果相对一般；而封装型特征选择算法则需要依赖在特征选择过程中的某种或多种学习算法，计算代价较过滤型算高，效率偏低，但特征选择效果相对较好。例如 Li 等人[3]结合遗传算法(Genetic Algorithm，GA)与 K 最近邻(k-Nearest Neighbor，kNN)分类器来选择特征子集；支持向量机(Support Vector Machine，SVM)建立结构风险最小化原则，具有很强的学习和泛化能力，例如文献[4-5]将二进制粒子群优化算法和支持向量机相结合，提出 BPSO-SVM 特征选择算法。此外，最小二乘支持向量机(LS-SVM)由 Suykens 等人提出，是标准支持向量机的一种改进[4]，通过求解一组线性方程来获取最优分类超平面，这有效避免了求解复杂的二次规划问题，大幅提升了求解速度，从而成为众多研究者解决不同领域问题时的有效工具。例如，Aydogdu 等人[18]将模糊聚类与 LS-SVM 整合，用于供水网络的故障率评估；Rocco Langone 等人[19]将 LS-SVM 用于工业设备的故障预测。

Pablo Moscato[7]首次提出 memetic algorithm 的概念。在标准遗传算法的流程中，通常对个体进行变异、交叉和选择，对个体的适应性进行迭代进化来得到问题的最优解[8]。然而对于文化基因算法，遗传操作的对象并不是种群空间中的普通个体。而是各局部区域推选出的优秀个体，遗传操作的目的是选出适应性强的优秀个体，同时也会通过交叉操作生成新的个体，这些新个体可能属于一些新的区域，在下一代局部搜索中它们会被邻近的优秀个体所取代，然后再进行新一轮的全局进化。基于全局搜索和局部搜索机制的文化基因算法在某些问题领域的搜索求解效率要比传统的遗传算法快几个数量级，应用领域较广并能得到满意的结果。例如 Ismail Karaoglan 等人[20]将文化基因算法用于混合载货问题的求解，Diego Cattaruzza 等人[21]将文化基因算法用于多车辆路径规划问题的研究。

本文研究封装型特征选择算法，为了降低计算的代价，选择最小二乘支持向量机 LS-SVM 作为特征选择方法的机器分类器，结合文化基因算法的搜索求解机制，提出基于 MA-LSSVM 的安全数据的高维特征选择算法。

2 相关工作

2.1 最小二乘支持向量机

最小二乘支持向量机 LS-SVM 将支持向量机中的不敏感损失函数变为二次损失函数，不等式约束变为等式约束，因此求解最优分类超平面的问题就转变成了一组线性方程组的求解，这能够有效避免求解复杂的二次规划问题，从而大幅提升求解速度。该问题描述如下[13-14]。

设有 l 个样本 $\{x_i,y_i\}_1^l$ 的训练集，其中，第 i 个输入数据 $x\in R^n$，第 i 个输出数据 $y_i\in\{-1,+1\}$ 是类别。可以用非线性映射 $\varphi(\cdot)$ 把样本从原空间映射到一个特征空间 $\varphi(x)=(\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_n))$。支持向量机模型的目的是构造一个如下形式的分类器：

$$f(x)=\text{sign}[w^{\mathrm{T}}\varphi(x)+b] \tag{1}$$

使得样本 x 能够被 $f(x)$ 正确分类，其中，w 为权向量，b 为阈值。

最小二乘支持向量机求解的优化问题如下：

$$\min_{w,e} J(w,e)=\frac{1}{2}w^{\mathrm{T}}\varphi(x)+b \tag{2}$$

其满足等式约束：

$$y_i[w^{\mathrm{T}}\varphi(x_i)+b]+e_i=1,i=1,2,\cdots,l \tag{3}$$

其对偶问题的 Lagrange 多项式为：

$$L(w,b,e,a)=J(w,e)-\sum_{i=1}^{l}\alpha_i\{y_i[w^{\mathrm{T}}\varphi(x_i)+b]-1+e_i\} \tag{4}$$

其中，$\alpha_i\in R$ 为 Lagrange 乘子，根据等式约束，其值正负皆可。最优化的条件是：

$$\begin{cases}\dfrac{\partial L}{\partial w}=0\rightarrow\sum_{i=1}^{l}\alpha_i y_i\varphi(x_i)\\ \dfrac{\partial L}{\partial b}=0\rightarrow-\sum_{i=1}^{l}\alpha_i y_i=0\\ \dfrac{\partial L}{\partial e_i}=0\rightarrow\alpha_i=\gamma e_i, & i=1,2,\cdots,l\\ \dfrac{\partial L}{\partial \alpha_i}=0\rightarrow y_i[w^{\mathrm{T}}\varphi(x_i)+b]-1+e_i, & i=1,2,\cdots,l\end{cases} \tag{5}$$

式(5)可以写成如下的线性方程组：

$$\begin{bmatrix}I&0&0&-Z^{\mathrm{T}}\\0&0&0&-y^{\mathrm{T}}\\0&0&\gamma I&-I\\Z&y&1&0\end{bmatrix}\begin{bmatrix}w\\b\\e\\\alpha\end{bmatrix}=\begin{bmatrix}0\\0\\0\\I\end{bmatrix} \tag{6}$$

其中，$Z=[\varphi(x_1)y_1,\varphi(x_2)y_2,\cdots,\varphi(x_l)y_l]^{\mathrm{T}}$，$e=[e_1,e_2,\cdots,e_l]^{\mathrm{T}}$，$y=[y_1,y_2,\cdots,y_l]$，$I=[1,\cdots,1]^{\mathrm{T}}$，$\alpha=[\alpha_1,\cdots,\alpha_l]^{\mathrm{T}}$。消去 e 和 w，再利用 Mercer 条件：

$$\Omega_{kj}=y_k y_j\varphi(x_k)^{\mathrm{T}}\varphi(x_j)=y_k y_j\psi(x_k,x_j),k,j=1,\cdots,l \tag{7}$$

得到的方程组与 b,α 有关。方程组(6)转化为：

$$\begin{bmatrix}0&-y^{\mathrm{T}}\\y&\Omega+\gamma^{-1}I\end{bmatrix}\begin{bmatrix}b\\\alpha\end{bmatrix}=\begin{bmatrix}0\\I\end{bmatrix} \tag{8}$$

设 $A=\Omega+\gamma^{-1}I$。因为 A 是一个对称半正定矩阵，所以 A^{-1} 存在。求解线性方程组式(8)得到的解如(9)所示：

$$b=\frac{y^{\mathrm{T}}A^{-1}I}{y^{\mathrm{T}}A^{-1}y}\alpha=A^{-1}(I-yb) \tag{9}$$

使用式(5)中的第一个等式替换式(1)中的 w，然后使用式(7)推导出式(10)：

$$f(x)=\text{sign}[\alpha_i y_i\psi(x,x_j)+b] \tag{10}$$

其中，α_i,b 是线性方程组式(8)的解，式(10)中的 $f(x)$ 就是所求的分类器。

2.2 稳定性的度量方法

对于特征选择研究来说，度量特征选择算法的稳定性是其中的一个关键点。特征选择结果的稳定性是通过评估不同特征选择结果的相似性来界定的，常用的度量方法是在有差异的训练样本集上进行特征选择，然后对得到的特征选择结果进行相似性比较，从而评估特征选择算法的稳定性。目前，根据特征选择结果展现方式的不同，有 3 种类型的相似性度量方法：权重法(Weighting)、排序法(Ranking)和子集法(Subset)[16]。

本文将使用以下几种常用的基于特征子集的相似性度量方法，对提出的 MA-LSSVM 特征选择算法的实验结果进行分析。

1)Jaccard 指数(Jaccard Index)

$$sim_{Jaccard}(f_i,f_j)=\frac{|f_i\cap f_j|}{|f_i\cup f_j|} \tag{11}$$

其中，f_i 为特征向量。Jaccard 系数的取值范围为[0,1]，0 表示两个特征选择结果的特征子集完全不同，1 表示完全相同。因此，Jaccard 系数越大，特征子集之间的相似性就越高。

2）Dice 系数（Dice's Coefficient）

$$sim_{Dice}(f_i,f_j)=\frac{2|f_i\cap f_j|}{|f_i|+|f_j|} \tag{12}$$

Dice 系数的取值范围为[0,1]，0 表示两个特征选择结果的特征子集完全不同，1 表示完全相同。因此，Dice 系数越大，特征子集之间的相似性就越高。一般而言，Dice 系数和 Jaccard 指数很类似，值得注意的是：Dice 系数在两个特征子集有交叉时能得出更准确的稳定性评估，而且能够处理不同大小的子集。

3）平均标准汉明距离（Average Normal Hamming Distance，ANHD）

$$sim_{ANHD}(f_i,f_j)=\frac{1}{m}\sum_{k=1}^{m}|f_i^k-f_j^k| \tag{13}$$

其中，m 为特征总数，f_i^k 表示 f_i 特征向量中第 k 维的值，其取值范围为{0,1}，0 表示该特征没有被选择，1 表示该特征被算法选择。平均标准汉明距离的取值范围为[0,1]，若取值越小，特征子集之间的相似性就越高。

4）Kuncheva 系数（Kuncheva Index，KI）

$$sim_{Kuncheva(f_i,f_j)}=\frac{|f_i\cap f_j|\cdot m-k^2}{k(m-k)} \tag{14}$$

随着所选特征子集基数的增加，它们之间重叠的可能性也会增加，Kuncheva 系数能够避免两个特征子集偶然的交叉重叠。Kuncheva 系数的取值范围为[−1,1]，其值越大，两个特征选择结果的特征子集就越相似。

3　基于文化基因算法的特征选择

给定一个特征集合 $F=\{f_1,f_2,\cdots,f_i,\cdots,f_n\}$，特征集合 F 的一个特征子集 F_S 能被表示为 $S=\{s_1,s_2,\cdots,s_i,\cdots,s_n\}$，$s_i\in\{0,1\}$，$i=1,2,\cdots,n$。其中，$s_i$ 表示特征集合中第 i 个特征 f_i 是否被选择，如果 f_i 被选择，则 $s_i=1$；否则 $s_i=0$。特征选择的目的是找出能使分类器效果达到最佳的特征子集，所以需要一个定量的标准来衡量特征子集的分类能力。在文化基因算法中，“个体”即为特征子集，适应度是评价特征子集对分类器性能影响的重要指标，也就是评价个体性能的指标。为此，提出的适应度函数应该包含两个标准：1）分类器验证的准确率。使用特征子集中的特征来训练分类器，用交叉验证结果来评价分类器的性能，通过此方式指导种群的进化。2）选择特征的数量。每个特征子集包含一定数量的特征，如果分类器在两个特征子集下的准确度相同，那么其中包含特征较少的子集就会被优先选择。在分类准确率和特征数量这两个因素中，需要重点考虑的是分类准确率。因此在文化基因算法中，对于种群的每个个体 F_s，其适应度函数确定为如下形式：

$$fit(F_i)=\beta*P(F_i)-\delta*N(F_i) \tag{15}$$

其中，$P(F_i)$是以从个体 F_i 中选出的特征构造分类器所获得的分类准确率，即以个体 F_i 中与“1”对应的特征作为特征子集，训练 LS-SVM 分类器得到的分类准确率。$N(F_i)$ 是个体 F_i 中包含“1”的数量（特征子集规模）。文化基因算法的目的在于寻找全局最大值，分类准确率越高，特征子集规模越小，代表适应度值越大，该特征子集就越有可能在个体的相互竞争中获胜。β 和 δ 是分类准确率和特征数量的权重参数，根据具体情况调整。β 值越大表示特征子集的分类准确率越受重视；同理，δ 值越大则表示特征数量越能获得重视。

准确地来说，文化基因算法是一种优化算法的框架，使用不同的搜索策略能够组成不同的文化基因算法。由于高维小样本数据特征数量多，训练样本少，本文考虑到特征选择问题的特点，采用二进制差异演化（Binary Differential Evolution，BDE）作为全局搜索策略[9]并将禁忌搜索（Tabu Search，TS）作为局部搜索策略[13]。图 1 示出了基于 MA-LSSVM 的安全数据的高维特征选择算法流程。

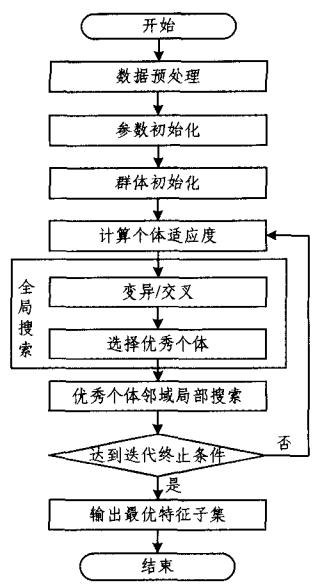

图 1　基于 MA-LSSVM 的安全数据的高维特征选择算法流程

其具体步骤描述如下：

步骤 1　数据预处理。为了方便进行数据分析，首先将样本集进行 z-score 标准化处理，把原始数据均转换为无量纲化的指标测评值，使各指标的值都处于同一个数量级。

步骤 2　参数初始化。在 BDE 算法中，需要设置初始种群大小 M、初始演化代数 $g=0$、最大迭代演化次数 g_{max}、缩放因子 ξ 和交叉因子 CR 的取值范围。在局部搜索的 TS 算法中，需要设置禁忌表长度 TL 和最大迭代步数 $Tstep_{max}$。

步骤 3　种群初始化。随机生成 M 个值在一定范围内的解决方案（个体），$g=1$。

步骤 4　根据适应度函数计算每个个体的适应度值。

步骤 5　全局搜索。根据 BDE 算法的优化机制，通过变异、交叉和选择，不断地更新个体的位置。

步骤 6　局部搜索。针对全局搜索中产生的优秀个体，使用 TS 算法对其邻域内的其他个体进行适应度值的计算，如果找到更优秀的个体，则更新最优解集。

步骤 7　判断终止条件。这里终止条件有两个：1）演化次数 g 超过最大允许演化次数 g_{max}；2）个体的适应度大于 0.99。如果达到这两个条件中的一个，文化基因算法的流程将会被终止，否则 $g=g+1$，跳转到步骤 4。

步骤 8　输出最优特征子集。

Tao Gong 针对离散优化问题对差分演化算法的变异操作进行了修改，提出了二进制差异演化（BDE），该算法在全局优化方面有着显著的效果[17]。在本文中，BDE 用来以确定最

优解为目的对整个解空间进行全局搜索，以及为局部搜索提供初始解。

合适的控制参数对于 BDE 算法来说十分重要，所以本文使用了一个自适应的控制参数调整策略。种群优化过程中，其适应度的方差为：

$$\sigma^2=\sum_{i=1}^{M}\left(\frac{fit(F_i)-fit_{avg}}{fit_{best}}\right) \tag{16}$$

其中，$fit(F_i)$为第 i 个个体的适应度值，fit_{avg} 为当前种群的平均适应度值，fit_{best} 为种群的最佳适应度值，M 为种群大小。

考虑种群的分布状况，BDE 算法的自适应的控制参数调整策略如下：

$$\xi^g=\xi_{\max}-(\xi_{\max}-\xi_{\min})\left(1-\frac{\sigma_g^2}{M}\right) \tag{17}$$

$$CR^g=CR_{\max}-(CR_{\max}-CR_{\min})\left(1-\frac{\sigma_g^2}{M}\right) \tag{18}$$

其中，g 代表了当前演化代，σ_g^2 代表第 g 代种群的适应度方差，ξ^g 代表当前进化代的缩放因子，$\xi_{\max}$ 与 $\xi_{\min}$ 分别代表缩放因子的最大值和最小值，CR^g 代表当前演化代的交叉因子，$CR_{\max}$ 与 $CR_{\min}$ 分别代表交叉因子的最大值和最小值。由此能够看出，σ^2 随着种群的进化而逐渐递减，同时缩放因子 ξ 逐渐减小，而交叉因子 CR 逐渐增大[10]。

局部搜索是文化基因算法中的一个重要概念，对算法的性能有着重要的影响。局部搜索的基本思想是基于贪婪算法在当前解决方案邻域不断寻找更优的解决方案，是一个推选局部区域内最优秀个体的过程，用来增强算法性能以及得到更好的个体适应度。TS 是一种启发式搜索的优化方法，拥有以下几个优点：高搜索效率、记忆功能和优秀的爬山能力。TS 算法通过禁忌表和相应的禁忌准则来避免迂回搜索，并使用藐视准则来赦免一些被禁忌的优良状态，然而其性能在很大程度上依赖于初始解[13]。

TS 算法作为局部搜索策略是为了在使用 BDE 算法做全局搜索时的每一个演化代上寻找更加优秀的个体，通过在局部邻域内的连续迭代，能够在局部邻域内找到最佳的解决方案。这种将 BDE 和 TS 优点结合在一起的搜索策略能够在解决方案和收敛速度上达到一个平衡。

4 实验和分析

本文相关实验在 PC 机上进行，该计算机的配置为 Intel Core2 Extreme Q6850 处理器，3.00GHz 频率，2GB 内存，Microsoft Windows 7，MATLAB 2012b +Weka 3.7.3。

实验数据使用 3 个高维小样本数据集 Colon，Leukemia 和 Lung 对算法性能进行验证，这些数据集均可从 BRB-AiTayTools 主页[1]下载。表 1 所列为这些数据集的相关信息，包括数据名称(Dataset)、特征总数(Features)、样本大小(Instances)以及类别数(Classes)。

表 1 实验数据集描述

Dataset	Features	Instances	Classes
Colon	2000	62	2
Leukemia	7129	72	2
Lung	12533	181	2

将本文提出的 MA-LSSVM 算法与 GA-kNN[3]，BPSO-SVM[4]两种封装式算法就特征子集生成的时间效率、特征选择的稳定性以及特征子集在分类器上的分类效果等指标进行比较。

设置种群大小 $M=100$，最大迭代演化次数 $g_{\max}=10000$，式(10)描述的 LS-SVM 算法的核函数 $\psi(\cdot)$采用径向基核函数 RBF，式(15)中 $\beta=100$，$\delta=0.5$，全局搜索中，初始缩放因子 $\xi=0.6$，交叉运算使用双交换，初始交叉因子 $CR=0.3$，局部搜索中，设置禁忌表长度 $TL=20$ 和最大迭代步数 $Tstep_{\max}=80$。然后采用十折交叉验证(10-fold cross-validation)来测试分类精度，即将数据集随机分成 10 份，轮流将其中 9 份用作训练集，余下 1 份用作测试集，结果的均值作为对算法精度的估计。为了得到更加精确的结果，需要进行多次 10 折交叉验证并求取均值，本文实验对每组算法进行 30 次 10 折交叉验证。本文提出的 MA-LSSVM 算法和对比算法的平均运行时间如表 2 所列。3 种算法的特征选择结果在随机森林(Rondom Forest，RF)和决策树 C4.5 两种分类器下的平均分类精度如表 3 所列，两种分类器都集成于 weka 中，其中，C4.5 中用于剪枝的置信因子 *confidenceFactor* 被设置为 0.25，Random Forest 的生成树的个数 *numTrees* 被设置为 10。

表 2 MA-LSSVM，GA-kNN 和 BPSO-SVM 3 种特征选择方法消耗的时间/s

Dataset	GA-kNN	BPSO-SVM	MA-LSSVM
Colon	31.55	27.43	18.32
Leukemia	67.43	50.98	36.08
Lung	224.02	198.82	160.91

表 3 MA-LSSVM，GA-kNN 和 BPSO-SVM 在数据集上的分类验证结果

Dataset	Used method	Feature subset size	C4.5/%	RF/%
Colon	MA-LSSVM	4	90.93	91.20
	GA-kNN	9	88.52	89.13
	BPSO-SVM	7	89.67	90.98
Leukemia	MA-LSSVM	14	98.63	99.01
	GA-kNN	15	94.12	95.33
	BPSO-SVM	15	95.53	95.90
Lung	MA-LSSVM	7	99.62	99.77
	GA-kNN	8	98.52	98.42
	BPSO-SVM	8	98.25	98.74

根据表 2 可知，MA-LSSVM 进行特征选择达到终止迭代的速度要明显快于 GA-kNN 和 BPSO-SVM 特征选择方法，结合表 3 可知，从被选择的特征选择子集中的特征个数以及特征子集在 C4.5 和 RF 两个分类器上的分类精度来看，MA-LSSVM 在 3 个数据集上都略优于 GA-kNN 和 BPSO-SVM。这说明 MA-LSSVM 在将全局搜索和局部搜索相结合的寻优过程中能够更有效地找到最优解，同时 LS-SVM 算法在处理非线性高维数据时的优异分类效果和速度也加快了种群的迭代演化过程，在获取特征子集的质量和速度两方面都得到提高。此外，表 4 使用了 4 种相似性度量方法来度量特征选择结果的稳定性，因为每种方法要对每个数据集进行 30 次实验，所以每种方法要对每个数据集计算 $30*(30-1)/2$ 次特征选择结果的相似性度量，表 4 中数据皆为各度量标准的平均值。通过 2.2 节的说明，证明了本文提出的 MA-LSSVM 算法的特征选择结果具有令人满意的稳定性。

1) http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html

表 4　MA-LSSVM，GA-kNN 和 BPSO-SVM 的稳定性比较

Dataset	Used method	Jaccard	Dice	ANHD	KI
Colon	MA-LSSVM	1.00	1.00	0.00	1.00
	GA-kNN	0.95	0.98	0.05	0.97
	BPSO-SVM	1.00	1.00	0.00	1.00
Leukemia	MA-LSSVM	1.00	1.00	0.00	1.00
	GA-kNN	0.90	0.95	0.03	0.90
	BPSO-SVM	1.00	1.00	0.00	1.00
Lung	MA-LSSVM	1.00	1.00	0.00	1.00
	GA-kNN	1.00	1.00	0.00	1.00
	BPSO-SVM	1.00	1.00	0.00	1.00

结束语　本文提出了一种基于文化基因算法和最小二乘支持向量机的安全数据特征处理方法(MA-LSSVM)。通过在 3 个高维样本数据集上与同类算法的实验对比，可得出 MA-LSSVM 算法在处理高维安全数据时具有较高的有效性和稳定性。本文提出的算法能取得较好效果的主要原因是：1)文化基因算法使用了基于 BDE 的全局搜索与基于 TS 的局部搜索相结合的优化策略框架，保证了算法能搜索到较好的解；2)全局搜索中控制参数的自适应调整加快了算法的收敛，避免陷入局部最优，提高了算法的性能；3)针对高维数据特征维度高、冗余多的特点，使用最小二乘支持向量机作为寻优过程中的分类器，提高了算法的整体执行效率。

参 考 文 献

[1] BELHUMEUR P N, HESPANHA J P, KRIEGMAN D. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720.

[2] SAEYS Y, INZA I, LARRAÑGA P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19): 2507-2517.

[3] LI L, WEINBERG C R, DARDEN T A, et al. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method[J]. Bioinformatics, 2001, 17(12): 1131-1142.

[4] LI K, GAO X, TIAN Z, et al. Using the curve moment and the PSO-SVM method to diagnose downhole conditions of a sucker rod pumping unit[J]. Petroleum Science, 2013, 10(1): 73-80.

[5] LI Y, ZHANG N, LI C. Support vector machine forecasting method improved by chaotic particle swarm optimization and its application[J]. Journal of Central South University of Technology, 2009, 16: 478-481.

[6] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.

[7] MOSCATO P. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms: C3P Report[R]. Caltech Concurrent Computation Program, 1989.

[8] ZHENG Y M. Improvement of feature selection method based on genetic algorithm [D]. Chongqing: Chongqing University, 2008. (in Chinese)
郑雅敏. 基于遗传算法的特征选择方法的改进研究[D]. 重庆：重庆大学，2008.

[9] WU Z F, HUANG H K, ZHAO X, et al. A Binary-Encoding Differential Evolution Algorithm for Agent Coalition[J]. Journal of Computer Research and Development, 2008, 45(5): 848-852. (in Chinese)
武志峰，黄厚宽，赵翔，等. 二进制编码差异演化算法在 Agent 联盟形成中的应用[J]. 计算机研究与发展，2008，45(5)：848-852.

[10] JIANG L Q, GUO Z, LIU G B, et al. Study on the strategy of scaling factor in differential evolution algorithm [C]// The first 2007 Advanced Instrumentation, Automation and Integration Technology Conference Proceedings. 2007: 508-510. (in Chinese)
姜立强，郭铮，刘光斌，等. 差分进化算法缩放因子取值策略研究[C]//2007 年首届仪表、自动化与先进集成技术大会论文汇编. 2007：508-510.

[11] LUO J, FAN P C. Improved particle swarm optimization based on genetic hybrid genes [J]. Computer Application Research, 2009, 26(10): 3716-3716, 3753. (in Chinese)
罗钧，樊鹏程. 基于遗传交叉因子的改进蜂群优化算法[J]. 计算机应用研究，2009，26(10)：3716-3717，3753.

[12] YANG W L, NING Y F. Integrated cross-factor and metropolis rule group search optimization algorithm [J]. Computer Engineering and Design, 2013, 34(6): 2020-2024. (in Chinese)
杨文璐，宁玉富. 基于交叉因子和模拟退火的群搜索优化算法[J]. 计算机工程与设计，2013，34(6)：2020-2024.

[13] SUN Y F. Hybrid strategy based on genetic algorithm and tabu search algorithm and its application [J]. Journal of Beijing University of Technology, 2006, 32(3): 258-262. (in Chinese)
孙艳丰. 基于遗传算法和禁忌搜索算法的混合策略及其应用[J]. 北京工业大学学报，2006，32(3)：258-262.

[14] LIU G, LI Y X, ZHENG H, et al. 2-Opt-and-generalized Opposition-based Differential Evolution Algorithm with Reserved Genes[J]. Journal of Chinese Computer Systmes, 2012, 33(4): 789-794. (in Chinese)
刘罡，李元香，郑昊，等. 保存基因的 2-Opt 一般反向差分演化算法[J]. 小型微型计算机系统，2012，33(4)：789-794.

[15] HE D K, WANG F L, ZHANG C M, et al. The genetic algorithm based on uniform design parameters [J]. Journal of Northeastern University (Natural Science Edition), 2003, 24(5): 409-411. (in Chinese)
何大阔，王福利，张春梅，等. 基于均匀设计的遗传算法参数设定[J]. 东北大学学报(自然科学版)，2003，24(5)：409-411.

[16] LIU Y. The stable feature selection research [J]. Micro Computer and Application, 2012, 31(15): 1-2. (in Chinese)
李云. 稳定的特征选择研究[J]. 微型机与应用，2012，31(15)：1-2.

[17] GONG T, TUSON A L. Differential evolution for binary encoding[M]// Soft Computing in Industrial Applications. Springer Berlin Heidelberg, 2007: 251-262.

[18] AYDOGDU M, FIRAT M. Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods[J]. Water Resources Management, 2015, 29(5): 1575-1590.

[19] LANGONE R, ALZATE C, DE KETELAERE B, et al. LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines[J]. Engineering Applications of Artificial Intelligence, 2015, 37(37): 268-278.

[20] KARAOGLAN I, ALTIPARMAK F. A memetic algorithm for the capacitated location-routing problem with mixed backhauls [J]. Computers & Operations Research, 2015, 55: 200-216.

[21] CATTARUZZA D, ABSI N, FEILLET D, et al. A memetic algorithm for the multi trip vehicle routing problem[J]. European Journal of Operational Research, 2014, 236(3): 833-848.