

基于 ABC-SVM 的考生行为自动识别^①

蔡丽霞¹, 马 琰²

¹(河南工业职业技术学院 网络管理中心, 南阳 473000)

²(河南工业职业技术学院 实验设备管理处, 南阳 473000)

摘 要: 针对支持向量机在考生行为自动识别中的参数优化问题, 提出了一种人工蜂群算法优化支持向量机的考生行为自动识别方法. 首先将支持向量机参数编码成为人工蜂群的蜜源, 以考生行为识别正确率作为搜索目标, 然后通过人工蜂群之间的信息交流和共享找到支持向量机的最优参数, 并建立最优考生行为识别模型, 最后采用仿真实验测试已建立考生行为识别模型的性能. 实验结果表明, 本文方法不仅提高了考生行为识别的正确率, 而且加快了考生行为识别的速度, 可以很好的满足考生行为自动识别实时性要求.

关键词: 考生行为识别; 支持向量机; 参数优化; 人工蜂群算法

Examinee Behavior Automatic Recognition Based on ABC-SVM

CAI Li-Xia¹, MA Yan²

¹(Network Management Center, Henan Polytechnic Institute, Nanyang 473000, China)

²(Laboratory and Equipment Management, Henan Polytechnic Institute, Nanyang 473000, China)

Abstract: According to the parameter optimization of support vector machine in the examinee behavior automatic recognition, an examinee behavior automatic recognition method based on artificial bee colony algorithm optimized parameters of support vector machine is proposed in this paper. Firstly, the parameters of support vector machine are encoded into artificial bee colony nectar and examinee behavior recognition correct rate is taken as searching target, and then the parameters of support vector machine is selected by exchange and sharing of information of artificial bee colony to establish the optimal examinee behavior recognition model, finally the performance is tested by simulation experiments. The experimental results show that, the proposed method not only improves the recognition correct rate of the examinee behavior, but also accelerate recognition speed, so it can meet the real-time requirements of examinee behavior recognition.

Key words: examinee behavior recognition; support vector machine; parameters optimization; artificial bee colony algorithm

在考试过程中, 存在着不同形式的作弊行为, 使考试很难公平、公正的进行, 尤其在大型考试中, 作弊行为给社会带来了极其恶劣的影响, 因此对考生作弊行为进行识别, 保证考试的公平性具有十分重要的意义^[1,2]. 随着计算机技术的发展, 考生行为自动识别已成为考试管理研究领域中的一个重要课题^[3].

由于考生行为自动识别具有十分重要的应用价值, 因此, 引起了国内外学者的广泛关注, 他们投入了大量的时间和精力进行了相关研究, 提出了许多有效的

考生行为自动识别方法^[4,5]. 由于考生行为的多样性, 给考生行为识别带来一定的难题, 传统考生行为识别方法易出现误识现象, 导致考生行为识别的正确率低^[6,7]. 近年来, 随着人工智能技术和机器学习算法的不断发现, 一些学者将它们引入到考生行为自动识别中, 出现了基于神经网络、支持向量机等考生行为识别方法^[8]. 神经网络是一种非线性分类能力强的机器学习算法, 但其要求样本数量大, 训练样本数量若不能满足“大样本”要求, 易出现“过拟合”现象, 同时存在

① 收稿时间:2014-08-25;收到修改稿时间:2014-10-20

神经元数目难以确定等不足^[9]. 支持向量机是一种针对小样本分类的机器学习算法, 克服了神经网络存在的不足, 可以得到更优的识别效果, 成为了当前考生行为识别的主要研究方向. 然而在实际应用中, SVM 算法的识别性能与其参数选择密切相关, 为了解决此难题, 学者们提出采用遗传算法、粒子群算法、模拟退火算法、蚁群算法等^[10-12]对 SVM 参数进行优化, 但这些算法均存在各自的不足, 影响了 SVM 的考生行为识别效果. 在考生行为识别研究中, 如何选择 SVM 参数, 提高考生行为自动识别的性能, 仍是一个未彻底解决的问题.

为了提高考生行为识别的正确率, 提出一种人工蜂群算法(artificial bee colony algorithm, ABC)优化支持向量机的考生行为自动识别方法(ABC-SVM). 首先将 SVM 参数编码为蜜源, 并以考生行为自动识别正确率作为目标函数, 然后通过模拟人工蜂群觅食过程找到 SVM 的最优参数, 最后根据最优参数建立考生行为自动识别模型, 并通过仿真实验测试 ABC-SVM 的性能. 仿真结果表明, ABC-SVM 不仅提高了考生行为识别的正确率, 降低了误识率, 而且加快了考生行为自动识别速度, 可以较好的满足考生行为识别实时性要求.

1 支持向量机和人工蜂群算法

1.1 支持向量机

支持向量机通过核函数 $\varphi(x)$ 将原空间的非线性问题映射到高维 Hilbert 特征空间, 从而将最优解问题转化为凸二次规划问题:

$$\begin{aligned} \min_{\omega, b, \xi} \quad & J(\omega, \xi) = \frac{1}{2} \omega^T \bullet \omega + C \sum_{i=1}^l \xi_i^2 \\ \text{s. t.} \quad & \begin{cases} y_i = \omega^T \bullet j(x_i) + b + \xi_i \\ \xi_i \geq 0 \\ i = 1, 2, L, 1 \end{cases} \end{aligned} \quad (1)$$

式中, ξ_i 为训练误差; 参数 C 为惩罚参数^[13].

引入对偶问题的 Lagrange 约束规划, 具体如下:

$$\begin{cases} L(\omega, b, \xi, \alpha) = J(\omega, \xi) - \sum_{i=1}^l \alpha_i (\omega \bullet \varphi(x_i) + b + \xi_i - y_i) \\ \alpha_i \geq 0 \\ i = 1, 2, L, 1 \end{cases} \quad (2)$$

式中, α_i 为 Lagrange 乘子.

根据极值定律, 对 Lagrange 式中的各参数 ω 、 b 、 ξ_i 、 α_i 求偏导, 可得:

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \\ \frac{\partial L}{\partial b} = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \\ \frac{\partial L}{\partial \alpha_i} = 0 \end{cases} \Rightarrow \begin{cases} \omega = \sum_{i=1}^l \alpha_i j(x_i) \\ \sum_{i=1}^l \alpha_i = 0 \\ \alpha_i = c \xi_i \\ \omega^T j(x_i) + b + \xi_i - y_i = 0 \end{cases} \quad (3)$$

式中, $i=1, 2, \dots, l$.

消去 ω 和 ξ_i , 得矩阵方程为:

$$\begin{bmatrix} 0 & I^T \\ I & j(x_i)^T j(x_i) + c^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \quad (4)$$

式中, $Z = [\varphi(x_1), \varphi(x_2), \dots, \varphi(x_l)]^T$, $Y = [y_1, y_2, \dots, y_l]^T$, $\rho = [I_1, I_2, \dots, I_l]^T$, $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_l]^T$, $\xi = [\xi_1, \xi_2, \dots, \xi_l]^T$, I 为单元矩阵.

根据 Mercer 条件, 可得:

$$\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j) \quad (5)$$

解上述方程组得:

$$\begin{cases} b = \frac{\rho^T (ZZ^T + c^{-1}I) Y}{\rho^T (ZZ^T + c^{-1}I) \rho} \\ \alpha = (ZZ^T + c^{-1}I)^{-1} (Y - b\rho) \end{cases} \quad (6)$$

支持向量机的分类函数为:

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (7)$$

由于 RBF 函数待优化参数少, 因此选择其建立支持向量机分类函数, 因此有:

$$f(x) = \text{sgn}(\alpha_i y_i \exp\left(-\frac{\|x - x_i\|}{\sigma}\right) + b) \quad (8)$$

式中, σ 为 RBF 核函数宽度.

为了分析参数 C 、 σ 不同值对 SVM 对分类性能的影响, 选择一个二分类问题, 选择 800 个样本作为训练集, 建立识别模型, 200 个样本作为测试集, 仿真结果具体见表 1, 表 2. 根据表 1 和 2 的结果可知, 参数 C 和 σ 取值不同, SVM 的分类结果不相同, 仿真结果表明, 要获得最优考生行为识别模型, 首先要获得最优 C 、 σ 值, 因此本文采用人工蜂群算法优化参数, 以提高考生行为识别的正确率.

表 1 $\sigma=1.65$ 时, 参数 C 对 SVM 分类结果的影响

C	分类正确率(%)
10	87.04
50	89.89
100	82.39
150	65.96
200	91.63
400	88.04
500	82.59
1000	89.63

表 2 $C=300$ 时, 参数 σ 对 SVM 分类结果的影响

σ	分类正确率(%)
0.125	76.94
0.25	67.05
0.5	74.05
1.0	65.02
1.5	88.74
2.0	82.33
5.0	94.28
10.0	89.57

1.2 人工蜂群算法

人工蜂群(ABC)算法是一种模拟蜂群觅食的智能算法, 食物源代表待优化问题的一个可能解, 蜂群采蜜的过程就是搜寻优化问题最优解的过程^[14]. 用 $x_i=(x_{i1}, x_{i2}, \dots, x_{iD})$ 表示第 i 个食物源, 首先随机产生 SN 个解(食物源), 引领蜂和跟随蜂根据式(9)在食物源的邻域生成一个候选食物源

$$v_{ij} = x_{ij} + r_{ij}(x_{ij} - x_{kj}) \quad (9)$$

式中, v_{ij} 是生成的候选食物源; r_{ij} 是 $[-1, 1]$ 上均匀分布的随机数, 它控制 x_{ij} 邻域的生成范围.

跟随蜂选择食物源的概率计算公式如下:

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (10)$$

在 ABC 算法中, 某个蜜蜂个体如果连续经过 limit 次循环之后食物源仍然没有得到更新, 个体就要放弃食物源, 转变为侦察蜂, 侦察蜂根据下式产生新的食物源.

$$x_i^j = x_{min}^j + rand1(x_{max}^j - x_{min}^j) \quad (11)$$

式中, x_i^j 为新的食物源的第 j 维分量; x_{min}^j 、 x_{max}^j 分别为第 j 维变量的最小值和最大值.

ABC 算法由于 r_{ij} 是一随机数, 使得该搜索策略具有较好的探索能力, 但却忽略了算法的开采能力. 为此, 有学者提出了一种新的搜索策略:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) + \varphi_{ij}(x_{bestj} - x_{ij}) \quad (12)$$

其中, ϕ_{ij} 和 φ_{ij} 均为随机数, x_{best} 为整个种群搜索到的最优位置.

由(12)式可以看出, 由于有最优位置 x_{best} 的引导, 上述搜索策略在一定程度上提高了算法的开采能力. 为进一步提高算法的开采能力, 受差分进化算法中不同变异操作模式的启发, 本文对 ABC 算法原有的搜索策略改进为:

$$v_{ij} = x_{bestj} + r_{ij}(x_{bestj} - x_{kj}) \quad (13)$$

由于式(13)有种群最优位置引导种群的搜索轨迹并且仅在最优位置附近产生新的候选解, 提高算法的开采能力.

为了测试本文对人工蜂群算法改进的有效性, 选用 2 个标准测试函数进行仿真实验, 测试函数具体为:

1) Sphere 函数

$$f(x) = \sum_{i=1}^n x_i^2 \quad (14)$$

2) Rosenbrock:

$$f_2(x) = \sum_{i=1}^{n-1} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \quad (15)$$

2 个测试函数求解曲线如图 1 和图 2 所示. 从图 1 和图 2 可知, 相对于标准人工蜂群算法, 改进人工蜂群算法收敛速度明显加快, 找到更优的解, 具有更加优异的性能.

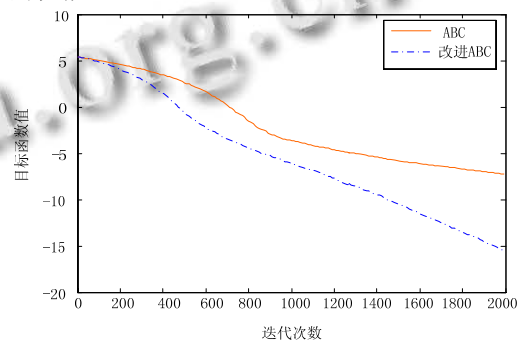


图 1 Sphere 函数进化曲线

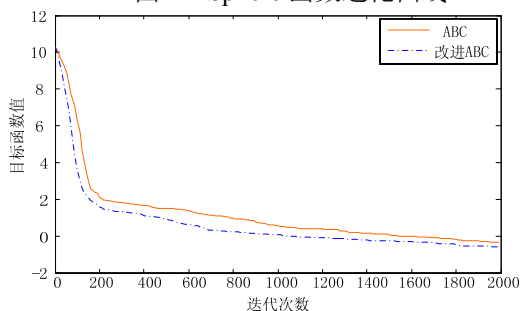


图 2 Rosenbrock 函数进化曲线

2 ABC-SVM的考生行为识别模型

2.1 SVM 参数优化的数学模型

对给定考生行为识别问题, SVM 的参数 C 、 σ 优化数学模型为:

$$\begin{aligned} & \max P(C, \sigma) \\ & s. t. \\ & \begin{cases} C \in [C_{min}, C_{max}] \\ \sigma \in [\sigma_{min}, \sigma_{max}] \end{cases} \end{aligned} \quad (16)$$

2.2 ABC 算法求解 SVM 参数的步骤

- 1) 初始化蜂群的大小(N_c), 采蜜蜂数量(N_e), 跟随蜂数量(N_o), 极限值($limit$), 最大循环次数(MCN).
- 2) 将 SVM 参数(C, σ)编码成初始解, 即采蜜蜂和食物源, 将收集的考生行为训练样本输入 SVM 进行学习, 计算考生行为识别的正确率, 并作其作为蜜源的适应函数.
- 3) 采蜜蜂根据式(9)从邻域搜索新的蜜源, 并计算其适应度值.
- 4) 跟随蜂根据式(10)计算出蜜源选择概率, 采用贪婪算法选择新的蜜源.

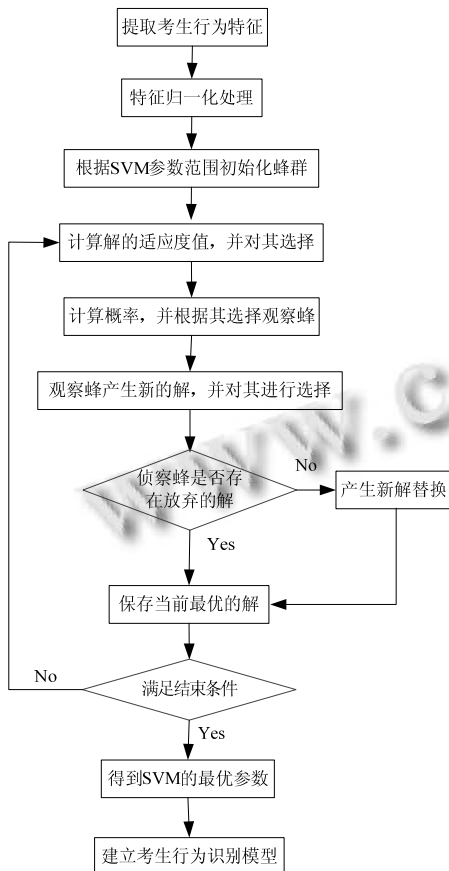


图3 ABC-SVM 的考生行为识别流程

5) 观察蜂根据采蜜蜂搜索的蜜源, 计算该蜜源的 P_i , 根据 P_i 从现有蜜源的邻域中搜索新蜜源进行选择.

6) 经过 $limit$ 次循环后, 判断是否有要丢掉的蜜源, 若存在, 则侦察蜂根据式(13)产生新解代替它.

7) 如果迭代次数大于最大循环次数 MCN , 则 SVM 参数寻优结束, 不然返回步骤 2)继续寻优.

8) 根据最优蜜源得到 SVM 的参数(C, σ), 建立考生行为识别模型.

9) 综合上述可知, 基于 ABC-SVM 的考生行为识别流程具体如图 3 所示.

3 仿真实验

3.1 数据来源

为了验证 ABC-SVM 的考生行为识别性能, 在 P4 双核心 2.8GHz CPU, 4GB RAM, Windows 7 操作系统的计算机上, 采用 Matlab 2012 进行仿真实验. 仿真数据来自某职业技术学院期末考试 5000 幅考生行为图像, 部分图像如图 4 所示. ABC 算法的参数设置为: 蜂群数量为 100, 侦察蜂的比例为 20%, 采蜜蜂数量为 20, 最大搜索次数设置为 30, 最大迭代次数为 200.

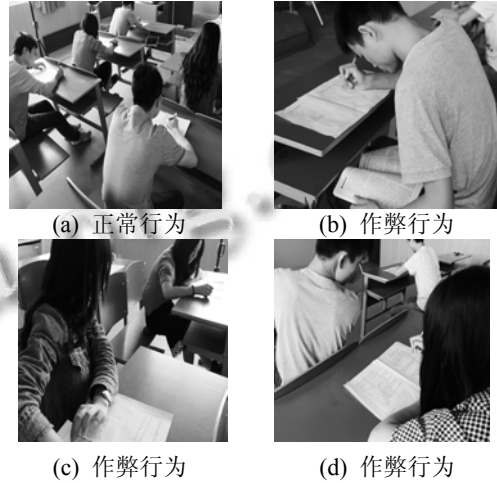


图4 部分考生行为图像

3.2 对比模型及性能评价标准

为了使 ABC-SVM 的识别结果更具说服力, 在相同条件下, 采用遗传算法优化 SVM(GA-SVM)和蚁群算优化 SVM(ACO-SVM)、采用粒子群算法优化 SVM(PSO-SVM)进行对比实验. 模型性能评价指标为: 识别正确率、误识率和平均识别时间, 识别正确率、误识率分别计算如下:

$$rate = \frac{\text{正确识别出的样本数}}{\text{样本总数}} \times 100\% \quad (17)$$

$$error = \frac{\text{错误识别样本数}}{\text{样本总数}} \times 100\% \quad (18)$$

3.3 结果与分析

3.3.1 识别结果对比

采用 GA、ACO、PSO、ABC 算法对 SVM 参数进行优化,得到的结果见表 3,然后采用表 3 的参数建立相应的考生行为识别模型,所有模型的考生行为识别结果如图 5~7。对图 5~7 的结果进行分析可知,相对于对比模型,ABC-SVM 的支持向量数相对较少,降低了计算杂度,提高了的考生行为识别正确率,这主要是由于 ABC 算法比 GA、ACO、PSO 算法的搜索能力更强,获得更优的 SVM 参数 C, σ, 加快了参数寻优的速度,有效降低了考生行为识别误识率,提高了考生行为识别正确率。

表 3 不同算法选择的支持向量机参数

SVM 参数优化算法	C	σ
GA	166.05	1.41
ACO	237.13	5.24
PSO	480.45	7.87
ABC	59.12	7.58

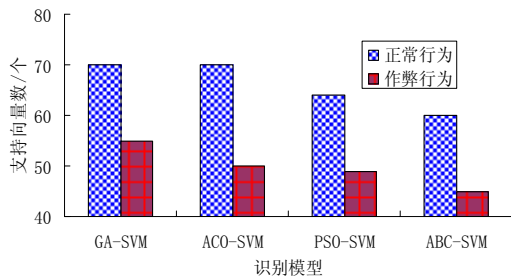


图 5 不同模型的支持向量数对比

3.3.2 考生行为识别速度对比

对于考生行为自动识别系统,考生行为识别速度十分关键,为此对不同模型平均识别时间(秒, s)进行仿真实验,所有模型的平均识别时间如图 8 所示。从图 8 可以看出,相对于对比模型,ABC-SVM 的平均识别时间最少,识别速度最快,从而提高了考生行为识别效率,这主要由于 ABC-SVM 大幅度减少了支持向量的数量,加快了算法的收敛速度,满足了考生行为识别的实时性要求。

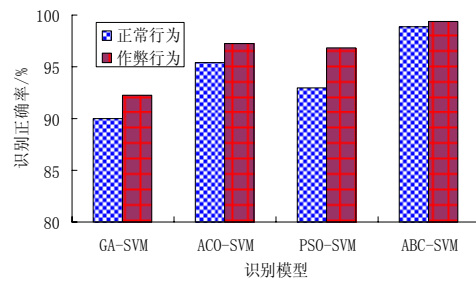


图 6 不同模型的考生行为识别正确率对比

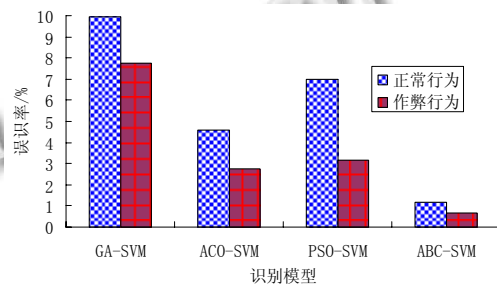


图 7 不同模型的考生行为误识率对比

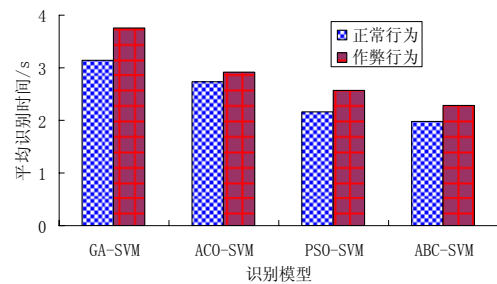


图 8 不同模型的平均识别时间对比

参考文献

- 1 吴芳,刘辉.视频监控考试系统在临床技能培训及考核中的应用.护理学杂志,2011,26(13):21-23.
- 2 李贵兵,谭颖,吴兵,杨杰.基于生物识别技术的考试身份认证系统的框架设计.计算机信息与技术,2013,12(5):5-9.
- 3 王超,魏启明,邓安远.无线指纹识别技术在考试系统的应用研究.计算机仿真,2010,27(1):309-313.
- 4 赵晴,李慧鹏.考试管理系统中的指纹识别与验证系统设计.河北工程技术高等专科学校学报,2008,6(2):37-40.
- 5 彭曙蓉,彭楚武,苏深广,杨文忠.指纹 IC 卡在考试身份认证系统中的应用.计算机与现代化,2003,98(10):85-87.
- 6 李仕强,王水平.基于指纹特征的考生身份认证系统.计算机工程,2006,32(21):160-162.
- 7 王尚平,王琪,张亚玲.基于身份认证的网络考试安全机制.

- 计算机工程,2009,35(18):136-138.
- 8 许敏,王士同.PSO 优化的神经网络在教学质量评价中的应用.计算机工程与设计,2008,29(20):5327-5330.
- 9 王洪斌,杨香兰,王洪瑞,等.一种改进的 RBF 神经网络学习算法.系统工程与电子技术,2002,24(6):103-105.
- 10 栾庆林,卢辉斌.自适应遗传算法优化神经网络的入侵检测研究.计算机工程与设计,2008,29(12):3022-3024.
- 11 Huang CL, Wang CJ. A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications, 2011, 31(2): 231-240.
- 12 Montemanni R, Smith DH, Gambardella LM. Ant colony systems for large sequential ordering problems. Proc. of the 2007 IEEE Swarm Intelligence Symposium, 2012, 12: 478-482.
- 13 邵信光,杨慧中,陈刚.基于粒子群优化算法的支持向量机参数选择及其应用.控制理论与应用,2006,23(5):740-743.
- 14 胡珂,李迅波,等.改进的人工蜂群算法性能.计算机应用,2011,31(4):1107-1110.

WWW.C-S-A.ORG.CN

WWW.C-S-A.ORG.CN