

基于CSAGA-LSSVM算法的坦克驾驶模拟训练数据分类挖掘

邓青^{1,2}, 薛青¹, 翟凯¹

(1. 陆军装甲兵学院 演训中心, 北京 100072; 2. 68303 部队, 青海 格尔木 816099)

摘要: 利用坦克驾驶模拟器进行训练是提高操作技能的重要方法。针对以往驾驶模拟训练采用统计分析方法难以从复杂训练数据中发现知识和规律的不足, 提出CSAGA-LSSVM算法对坦克驾驶模拟训练数据进行分析。首先选择关键点快速生成Shapelets, 以减少候选Shapelets数量; 其次, 根据距离和时间间隔对Shapelets进行组合, 增强特征辨识能力; 然后, 设计自适应遗传算法, 动态调整交叉、变异概率, 寻找最小二乘支持向量机最优参数解, 提高分类结果的准确性; 与其他分类方法进行实验对比, 验证了CSAGA-LSSVM算法的可行性与有效性。最后, 将算法应用于某型坦克驾驶模拟器换挡操作数据的分类挖掘, 提取不同训练水平人员的操作特征, 促进指导个性化训练。

关键词: 坦克驾驶模拟器; 支持向量机; Shapelets特征; 遗传算法; 分类挖掘

中图分类号: TP391.9

文献标志码: A

Data classification mining of tank driving simulation training based on CSAGA-LSSVM algorithm

DENG Qing^{1,2}, XUE Qing¹, ZHAI Kai¹

(1. Training Center, Academy of Army Armored Forces, Beijing 100072, China;

2. 68303 Troops, Geermu, Qinghai 816009, China)

Abstract: Training with tank driving simulators is an important method to improve operation skills. In view of the difficulty to find knowledge and rules from complex training data by statistical analysis method in the past driving simulation training, this paper proposed the CSAGA-LSSVM classification mining algorithm to analyze tank driving simulation training data. Firstly, key points were quickly selected to generate shapelets and reduce the number of candidate shapelets. Secondly, the shapelets were then grouped according to distance and time interval to enhance the ability of feature identification. An adaptive genetic algorithm was designed to dynamically adjust the probability of crossover and mutation to find the optimal parameter solution of least squares support vector machine and improve the accuracy of classification results. Compared with other classification methods, the feasibility and effectiveness of the CSAGA-LSSVM algorithm were further verified. Finally, the algorithm was applied to the classification mining of gear shift operation data from a certain tank driving simulator to extract the operation characteristics of personnel with different training levels and guide personalized training.

Key words: tank driving simulator; support vector machine; shapelets feature; genetic algorithm; classification mining

坦克驾驶模拟训练是装甲兵掌握驾驶技能的重要途径, 对提高装甲分队战场快速机动能力具有重要意义^[1-2]。坦克驾驶模拟训练数据包括训练操作数据、受训人员数据等, 这些数据之间蕴含着复杂的关系。传统的坦克驾驶模拟训练结果由人工进行统计分析^[3], 易受分析人员专业知识、个人偏好的主观影响, 对训练

收稿日期: 2020-08-05

基金项目: 军内科研基金项目(JY2019C095)

作者简介: 邓青(1985—), 男, 江西新建人, 博士, 主要从事驾驶模拟训练系统、数据挖掘研究。E-mail: 154247597@qq.com

薛青(1961—), 男, 辽宁锦州人, 教授, 博士, 主要从事模拟仿真、数据挖掘研究。

的影响因素考虑不全,无法精确指导受训人员,也难以从这些复杂的数据中发现有价值的训练规律。为解决这一问题,提出将分类挖掘引入坦克驾驶模拟训练数据分析中,以期从中获取训练指导规律。

Robinson 等^[4]运用主成分分析方法对通信指挥模拟训练数据进行分析,为实施作战指挥提供支持,但需要事先提取指挥员的能力特征表示,不同场景下的概念描述比较繁杂。Cady^[5]选择装甲装备模拟训练系统中的地形条件、任务样式、敌人属性等作为输入数据,运用定制的聚类算法和最小二乘法进行分析,得出杀伤力、战损率等结果与输入数据之间的关系,但算法细节没有描述。Wang 等^[6]采用聚类方法分析装甲兵模拟训练数据,通过发现数据分布的簇特征,去除离群值,计算车辆位置的平均值,得出中心点移动速度与各车速度之间的关系,为机动力评估提供依据,但均值计算存在有偏估计问题。邓桂龙等^[7]运用关联规则方法分析某型空地作战系统模拟训练数据,提取雷达连续照射时间、打击导弹阵地种类与作战效果之间的强关联规则,辅助作战分析人员获取更多有用的知识,但生成规则的数量庞大,需要人工筛选。唐志武等^[8]采集了装甲兵模拟训练系统在特殊想定条件下的交战数据,运用多变量决策树进行分析,得到机动、防护等作战能力指标的影响因素,但决策树的分叉比较繁琐。贝叶斯网络运用有向图模型来描述和计算变量之间的概率依赖关系,利用先验知识更新后验概率,但往往需要专家确定初始值^[9],而坦克驾驶模拟训练数据在很多情况下无法提前确定先验概率。决策树分类具有原理简单、抗噪音强等优点^[10],主要适用于小样本数据集的分类,在面对海量高维度数据时易产生无效节点。深度学习在图像识别、自然语言理解领域有着广泛运用^[11],适合从高维度数据中提取特征,但由于内部的“黑盒”,对提取特征可解释性较差。最小二乘支持向量机(least squares support vector machine, LSSVM)利用核函数、特征空间等处理高维数据,实现从样本空间到特征空间的映射,具有良好的分类能力^[12-13]。

坦克驾驶模拟训练数据具有时序性的特点,不能直接输入 LSSVM 进行分类,而要提取相应的特征后进行分类,同时 LSSVM 存在的超参选择难问题也需要解决。本研究在 LSSVM 中引入 Shapelets 并进行组合以提取原始数据特征,设计自适应遗传算法(adaptive genetic algorithm, AGA)实现超参数的最优选择,最终得到一个组合 Shapelets 的自适应遗传最小二乘支持向量机算法(combination shapelets adaptive genetic algorithm-least squares support vector machine, CSAGA-LSSVM),将其应用于坦克驾驶模拟训练操作数据分析,通过分类得到标准的驾驶操作动作集合,更好地对受训人员进行指导。

1 坦克驾驶模拟训练数据分类挖掘问题描述

坦克驾驶模拟训练数据中的知识和信息是分析决策的有效依据。对这些数据进行分类挖掘,除了具有传统意义上的分类挖掘含义外,更重要的目标是从中发现有意义的模式。

从模拟器的数据采集分系统中抽取数据集 $B = (B_1, B_2, \dots, B_n)$, 将 B 划分为若干个不重复的记录,其中每个元组都对应相同数量的条件属性和类别属性,假设条件属性值的集合为 $C = \{C_1, C_2, \dots, C_n\}$, 类别属性值的集合为 $D = \{D_1, D_2, \dots, D_n\}$, 则分类挖掘问题为发现从 C 到 D 的映射 $f: C \rightarrow D$ ^[14]。在此基础上,将坦克驾驶模拟训练数据分类挖掘描述如下:

假设 $P = \{P^{(k)} \mid k=1, 2, \dots, N\}$ 为坦克驾驶模拟训练数据的 N 维特征空间 $U \subset \mathbf{R}^N$ 中所包含的非空模式集,其中任一子集族记为 $S = \{S^{(i)} \mid i=1, 2, \dots, c, 1 \leq c \leq N\}$, 若约束条件

$$\begin{cases} S^{(i)} \neq \emptyset, \\ S^{(i)} \cap S^{(j)} = \emptyset, \quad i \neq j, \\ P = \bigcup_{i=1}^c S^{(i)} \end{cases} \quad (1)$$

成立,则称 S 是对 N 维特征空间划分所形成的一个子类,即 S 为 P 中的一个分类。

2 CSAGA-LSSVM 分类挖掘算法

$TS = \{vt_1, vt_2, \dots, vt_i, \dots, vt_n\}$ 是一个实值时间序列,其中 vt_i 为时间序列点, n 为时间序列长度;记为

$|TS|$ 。通常相邻时间序列点之间是等间隔的,时间序列简记为 $TS = \{v_1, v_2, \dots, v_i, \dots, v_n\}$ 。时间序列的子序列 $TS_{j,l} = \{vt_j, vt_{j+1}, \dots, vt_{j+l-1}\}$ 是一个从 TS 取出起始位置为 j 、长度为 l 的连续序列。 m 个时间序列 TS 组成的集合称作时间序列数据集,记为 $DTS = \{TS_1, TS_2, \dots, TS_i, \dots, TS_m\}$, 其中时间序列数据集的实例数 $|DTS| = m$ 。

Shapelets 是时间序列 TS 中能够最大程度确定该序列所属类标签的特征表示,也是最具有辨识性和可解释性的局部时序模式。一个长度为 k 的 Shapelets $= \{s_1, s_2, \dots, s_k\}$ 是时间序列 $TS = \{v_1, v_2, \dots, v_i, \dots, v_n\}$ 的一个子序列。在时间序列分类问题中,从时间序列数据 DTS 集中每次学习获得的时间序列 Shapelets 长度并不相同。

2.1 基于关键点的 Shapelets 快速获取

Shapelets 获取是整个分类算法的第一步,通过定义关键点筛选子序列,减少产生的候选 Shapelets 个数,从而快速获取 Shapelets^[15]。

关键点是从时间序列点内部产生的,并能够表达时间序列的主要特征和变化趋势。本研究的关键点包括时间序列的起始点、结束点、阶跃点和极值点。

起始点、结束点位于时间序列的两端,分别表示整个时间序列的开始与结束时刻。在进行坦克驾驶模拟训练时,用来表示完成某一动作所用的时间。阶跃点 v_i^{step} 是指某一点与其相邻两点所构成线段的斜率之差超过预设阈值 ρ , 即

$$(v_{i+1} - v_i^{\text{step}}) / (v_i^{\text{step}} - v_{i-1}) \geq \rho \quad (2)$$

时(如坦克驾驶模拟器在完成发动起车后,松开离合器踏板要做到前 2/3 快、后 1/3 慢),对应的中间操作转换时刻被视为阶跃点。极值点包括局部极大值点 v_i^{lmax} 和极小值点 v_i^{lmin} , 分别满足:

$$\begin{aligned} v_{i-k} &\leq v_i^{\text{lmax}} \text{ 且 } v_{i+k} \leq v_i^{\text{lmax}}, \\ v_i^{\text{lmin}} &\leq v_{i-k} \text{ 且 } v_i^{\text{lmin}} \leq v_{i+k}. \end{aligned} \quad (3)$$

其中 k 是极值点附近邻域的大小。

确定关键点后,按以下步骤产生 Shapelets:

- 1) Key 为存储关键点的数组, $Dsup$ 为存储 Shapelets 的字典,初始化均为空。设定初始邻域 k 、阈值 ρ 。 $TS = \{v_1, v_2, \dots, v_n\}$ 为一时间序列, n 为时间序列长度, $i = 1$;
- 2) 判断点 v_i 是否为时间序列的起始点,若是则将点 v_i 存入数组 Key 中,转步骤 6);
- 3) 判断点 v_i 是否满足 $v_{i-k} \leq v_i$ 且 $v_{i+k} \leq v_i$ 或 $v_i \leq v_{i-k}$ 且 $v_i \leq v_{i+k}$, 若是则将点 v_i 存入数组 Key 中,转步骤 6);
- 4) 判断点 v_i 是否满足 $(v_{i+1} - v_i) / (v_i - v_{i-1}) \geq \rho$, 若是则将点 v_i 存入数组 Key 中,转步骤 6);
- 5) 判断点 v_i 是否为时间序列的结束点,若是则将点 v_i 存入数组 Key 中,转步骤 7);
- 6) 结束对点 v_i 的处理, $i = i + 1$, 转步骤 2);
- 7) 对关键点数组 Key 进行遍历,按顺序依次从 Key 中取出两个元素构成 Shapelets 的两个端点,并将该 Shapelets 及其所在的序号存入 $Dsup$ 。

2.2 组合 Shapelets 特征生成

通过坦克驾驶模拟器可以进行各种驾驶操作练习,对油门、离合等部件的操作有明确的先后次序和时间间隔要求,因此受训人员的驾驶动作识别属于多维时间序列分类问题。而之前获取的 Shapelets 由于忽略不同序列之间的操作时间、逻辑组合等关系,难以达到准确的辨识效果。因此将多个 Shapelets 进行组合,并加入时间间隔,以增强组合 Shapelets 的分类能力,并通过信息增益评价最优 Shapelets 组合,以提高分类的准确性。

对于长度均为 n 的 TS_1 和 TS_2 , 采用欧氏距离

$$dis(TS_1, TS_2) = \sqrt{\sum_{i=1}^n (TS_1(i) - TS_2(i))^2} \quad (4)$$

作为两个时间序列之间的距离,来度量两者间的差异。

对于 Shapelets 子序列 s 与 TS_i 的距离 $shpdis(s, TS_i)$, 采用滑动窗口方式在 TS_i 上生成 $(n - s + 1)$ 个

与 s 等长的序列 $TS_{i,|s|}$, 按式(4)计算距离, 由动态弯曲距离比对原则取最小值作为 s 与 TS_1 之间的距离度量, 即:

$$shpdis(s, TS_i) = \min(dis(s, TS_{i,|s|}))。 \quad (5)$$

对于时间序列数据集 $DTS = \{TS_1, TS_2, \dots, TS_i, \dots, TS_m\}$, 类标签的个数为 C , 设某一类 c_i 在时间序列数据集 DTS 中有 n_i 个时间序列, 则时间序列数据集 DTS 的熵表示为:

$$E(DTS) = - \sum_{i=1}^C \frac{n_i}{m} \log_2 \frac{n_i}{m}。 \quad (6)$$

随机选择 s_1, s_2 两个 Shapelets 进行合取操作, 按式(5)分别计算 s_1, s_2 与 $DTS = \{TS_1, TS_2, \dots, TS_i, \dots, TS_m\}$ 中每一个时间序列之间的距离。合取后的 $s_1 \wedge s_2$ 同时含有 s_1, s_2 的特征, 则 $s_1 \wedge s_2$ 与 TS_i 的距离 $comshpdis(s_1 \wedge s_2, TS_i)$ 应取两者中较大的值进行度量:

$$comshpdis(s_1 \wedge s_2, TS_i) = \max(shpdis(s_1, TS_i), shpdis(s_2, TS_i))。 \quad (7)$$

对 $s_1 \wedge s_2$ 与 $DTS = \{TS_1, TS_2, \dots, TS_i, \dots, TS_m\}$ 中时间序列的距离按递增顺序存入一维数组 $Discshp$ 。设定阈值 δ , 遍历数组 $Discshp$ 元素并与 δ 比较, 将时间序列数据集 DTS 分成 DTS_{left}, DTS_{right} 两部分:

$$DTS_{left} = \{TS_i \mid TS_i \in DTS, Discshp[i] \leq \delta\}, \quad (8)$$

$$DTS_{right} = \{TS_i \mid TS_i \in DTS, Discshp[i] > \delta\}。 \quad (9)$$

通过式(6)计算 DTS_{left}, DTS_{right} 数据集的熵, 得到 $s_1 \wedge s_2$ 的信息增益

$$I(s_1 \wedge s_2, \delta) = E(DTS) - \frac{|DTS_{left}|}{m} E(DTS_{left}) - \frac{|DTS_{right}|}{m} E(DTS_{right}) \quad (10)$$

以对 $s_1 \wedge s_2$ 进行评估, 并将对应的子序列加入候选组合 Shapelets, 记作 $Canshp$ 。

为增强组合 Shapelets 的分类能力, 参照决策树的构建思想, 将 Shapelets 之间的时间间隔作为分类的辅助特征, 减少单纯使用距离度量分为两个数据集 DTS_{left}, DTS_{right} 所产生的分类误差。在式(5)中, 通过求取 s 与 TS_i 的距离最小值, 找到最佳匹配初始位置记作 $Bpos(s, TS_i)$, $TS_i(Bpos(s, TS_i), Bpos(s, TS_i) + s - 1)$ 表示时间序列 TS_i 从初始位置 $Bpos(s, TS_i)$ 开始, 长度为 $|s|$ 的子序列, 则有:

$$shpdis(s, TS_i) = dis(s, TS_i(Bpos(s, TS_i), Bpos(s, TS_i) + s - 1))。 \quad (11)$$

设 $Bpos(s_1, TS_i), Bpos(s_2, TS_i)$ 分别表示 s_1, s_2 与 TS_i 的最佳匹配初始位置, 并且 $Bpos(s_1, TS_i) < Bpos(s_2, TS_i)$, 则组合后的 $s_1 \wedge s_2$ 在时间序列数据集上的时间间隔

$$Itime(s_1 \wedge s_2) = Bpos(s_2, TS_i) - Bpos(s_1, TS_i)。 \quad (12)$$

对 $Canshp$ 中的所有组合 Shapelets 按式(12)计算相应子序列的时间间隔, 得到时间间隔集合

$$SItime = \{Itime(s_i \wedge s_j) \mid Itime(s_i \wedge s_j) = Bpos(s_j, TS) - Bpos(s_i, TS), s_i, s_j \in Canshp\}。 \quad (13)$$

设定时间间隔阈值 θ , 将 $SItime$ 与 θ 进行差值比较, 将时间序列数据集 DTS_{left}, DTS_{right} 进行分割, 以 DTS_{left} 为例, 可形成 $DTS_{left}^1, DTS_{left}^2$ 两部分数据集:

$$DTS_{left}^1 = \{TS_i \mid TS_i \in DTS_{left}, SItime[i] \leq \theta\}, \quad (14)$$

$$DTS_{left}^2 = \{TS_i \mid TS_i \in DTS_{left}, SItime[i] > \theta\}。 \quad (15)$$

通过式(6)、(10)计算得到信息增益

$$I(s_1 \wedge s_2, \theta) = E(DTS_{left}) - \frac{|DTS_{left}^1|}{|DTS_{left}|} E(DTS_{left}^1) - \frac{|DTS_{left}^2|}{|DTS_{left}|} E(DTS_{left}^2), \quad (16)$$

选择信息增益最大的子序列作为组合 Shapelets。

2.3 基于自适应遗传算法的 LSSVM 超参数寻优

考虑到用 LSSVM 分类时, 核参数 σ 和惩罚因子 C 的取值对分类算法的性能有重要影响, 而且遗传算法具有良好的鲁棒性, 故本节采用遗传算法进行参数寻优。但传统的遗传算法存在早熟收敛等问题, 故采用自适应遗传算法改进交叉、变异操作, 避免陷入局部最优。具体步骤如下。

1) 交叉操作

采用单点交叉算子, 随机选择两个父代个体, 并设置每一对个体的基因位置用于交叉操作。随着进化代

数的增加,按式(17)自动调整交叉概率 p_c ,然后在交叉点进行染色体互换,生成两个新的子个体。这种自适应变化克服了因采用固定交叉概率而导致的后期不易保留优秀个体的问题,有利于维护种群的多样性。

$$p_c = \frac{\exp(-0.5\tau)}{p_{size} \cdot \sqrt{len}}, \quad (17)$$

式中: τ 为进化代数, p_{size} 为种群数量, len 为染色体长度。

2) 变异操作

变异操作促使产生新个体,可以提高算法的局部搜索能力,变异概率设定过高容易破坏适应度最优的个体,导致收敛变慢;过低不利于产生新个体,容易陷入早熟。根据个体的适应度变化动态调整变异概率 p_m ,对于适应度较小的个体赋予较大的变异概率,促进个体向更优解进化,反之赋予较小概率,即:

$$p_m = \begin{cases} p_{m0} - \frac{\lambda (fit_{\max} - fit_i)}{fit_{\max} - fit_{\text{avg}}}, & f_i \geq f_{\text{avg}}; \\ p_{m0} + \frac{\lambda (fit_{\text{avg}} - fit_i)}{fit_{\text{avg}} - fit_{\min}}, & f_i < f_{\text{avg}}. \end{cases} \quad (18)$$

式中: fit_{\max} 为种群中的最大适应度值, fit_{\min} 为种群中的最小适应度值, fit_i 为种群中个体 i 的适应度值, λ 为调整系数, p_{m0} 为初始概率常数。

2.4 基于组合 Shapelets 的 AGA-LSSVM 分类

在获得组合 Shapelets 的基础上,利用 Shapelets 转换概念^[16-17]计算原始时间序列与 Shapelets 之间的距离,将原始序列映射到新的特征空间,通过 AGA-LSSVM 算法进行分类。

假设 $\mathbf{x}_k \in \mathbf{R}^n$ 为经转换后的输入, $y_k \in \mathbf{R}$ 为类标签,则 LSSVM 将分类转化为如下优化问题:

$$\min \phi(\boldsymbol{\omega}, \mathbf{e}) = \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} + \frac{1}{2} C \sum_{k=1}^N e_k^2, \text{ s.t. } y_k [\boldsymbol{\omega}^T \boldsymbol{\varphi}(\mathbf{x}_k) + b] = 1 - e_k. \quad (19)$$

式中: $\boldsymbol{\omega}$ 为权重矩阵, b 为偏倚变量, e_k 为训练误差, C 为惩罚因子。

对式(19)的优化问题采用拉格朗日泛函得到下式:

$$L(\boldsymbol{\omega}, b, \mathbf{e}, \boldsymbol{\alpha}) = \phi(\boldsymbol{\omega}, \mathbf{e}) - \sum_{k=1}^N \alpha_k (y_k [\boldsymbol{\omega}^T \boldsymbol{\varphi}(\mathbf{x}_k) + b] - 1 + e_k). \quad (20)$$

式中, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_k, \dots, \alpha_N]^T$ 为拉格朗日乘子矩阵。根据鞍点条件可得:

$$\begin{cases} \frac{\partial L}{\partial \boldsymbol{\omega}} = 0 \rightarrow \boldsymbol{\omega} - \sum_{k=1}^N \alpha_k \boldsymbol{\varphi}(\mathbf{x}_k) = 0, \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^N \alpha_k y_k = 0, \\ \frac{\partial L}{\partial e_k} = 0 \rightarrow y_k [\boldsymbol{\omega}^T \boldsymbol{\varphi}(\mathbf{x}_k) + b] - 1 + e_k = 0, \\ \frac{\partial L}{\partial \alpha_k} = 0 \rightarrow \alpha_k - C e_k = 0. \end{cases} \quad (21)$$

结合卡库塔条件(Karush-Kuhn-Tucker, KKT),将式(21)转换为矩阵形式所表示的线性方程组:

$$\begin{bmatrix} 0 & \mathbf{H}^T \\ \mathbf{H} & \boldsymbol{\Omega} + C^{-1} \mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Y} \end{bmatrix}, \quad (22)$$

式中, $\mathbf{H} = [1, 1, \dots, 1]_{N \times 1}^T$, \mathbf{I} 为 N 阶单位矩阵, $\mathbf{Y} = [y_1, y_2, \dots, y_N]^T$,

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\varphi}(\mathbf{x}_1) \cdot \boldsymbol{\varphi}(\mathbf{x}_1) & \boldsymbol{\varphi}(\mathbf{x}_1) \cdot \boldsymbol{\varphi}(\mathbf{x}_2) & \cdots & \boldsymbol{\varphi}(\mathbf{x}_1) \cdot \boldsymbol{\varphi}(\mathbf{x}_N) \\ \boldsymbol{\varphi}(\mathbf{x}_2) \cdot \boldsymbol{\varphi}(\mathbf{x}_1) & \boldsymbol{\varphi}(\mathbf{x}_2) \cdot \boldsymbol{\varphi}(\mathbf{x}_2) & \cdots & \boldsymbol{\varphi}(\mathbf{x}_2) \cdot \boldsymbol{\varphi}(\mathbf{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{\varphi}(\mathbf{x}_N) \cdot \boldsymbol{\varphi}(\mathbf{x}_1) & \boldsymbol{\varphi}(\mathbf{x}_N) \cdot \boldsymbol{\varphi}(\mathbf{x}_2) & \cdots & \boldsymbol{\varphi}(\mathbf{x}_N) \cdot \boldsymbol{\varphi}(\mathbf{x}_N) \end{bmatrix}, \quad k(x_i, x_j) = \boldsymbol{\varphi}(x_i) \cdot \boldsymbol{\varphi}(x_j) \text{ 为核函数。}$$

令 $\mathbf{A} = \mathbf{\Omega} + \mathbf{C}^{-1} \mathbf{I}$, 最终求得 α 和 b 分别为:

$$\alpha = \mathbf{A}^{-1} (\mathbf{Y} - b \mathbf{H}) , \quad (23)$$

$$b = \frac{\mathbf{H}^T \mathbf{A}^{-1} \mathbf{Y}}{\mathbf{H}^T \mathbf{A}^{-1} \mathbf{H}} . \quad (24)$$

根据 α 和 b 的值, 对任意输入样本 \mathbf{x} 的分类函数为:

$$\hat{y} = \sum_{j=1}^N \alpha_j y_j k(\mathbf{x}_i, \mathbf{x}_j) + b . \quad (25)$$

对于多维时间序列数据集 $DTS = \{TS_1, TS_2, \dots, TS_i, \dots, TS_m\}$ 的任一维 TS_i , 其输入样本子集可采用 AGA-LSSVM 方法学习训练出分类器, 根据符号函数 $\text{sgn}()$ 计算输出类标签, 为:

$$f(\mathbf{x})_i = \text{sgn} \left(\sum_{j=1}^N \alpha_j y_j k(\mathbf{x}_i, \mathbf{x}_j) + b \right) . \quad (26)$$

通过以上对 CSAGA-LSSVM 分类挖掘算法四个阶段的描述, 得到 CSAGA-LSSVM 算法具体流程如图 1 所示。

3 实验与对比分析

为了验证 CSAGA-LSSVM 分类挖掘算法性能, 设计多维时间序列数据的分类实验, 分析该算法的分类精度和执行效率。

实验用计算机配置: 处理器 Intel(R) Core(TM) i7-4510@2.60 GHz, 硬盘 1 TB+128 GB SSD, 显卡 NVIDIA GeForce 840 M, 内存 8 GB DDR3, 操作系统 Windows 7 64 位, 编程环境为 Python 3.7。选择公开数据集 Trace、ItalyPower、Wafer 和 Coffee 进行对比实验, 数据集的简要说明如表 1 所示。

表 1 实验数据集
Tab. 1 Experiment data set

数据集名	数据对象总数	序列长度	类标签数
Trace	200	275	4
ItalyPower	1 096	24	2
Wafer	1 194	198	2
Coffee	56	286	2

实验中分别利用基于采样 Shapelets^[18] (S-Shapelets)、逻辑 Shapelets^[19] (L-Shapelets) 和本研究的组

合 Shapelets (C-Shapelets) 对原始数据进行特征提取, 其中 S-Shapelets 通过划分子类得到中心序列, 选择与中心序列距离之和最小的序列作为采样数据, 从而减少候选 Shapelets 的产生; L-Shapelets 采用加速技术与剪枝方法过滤具有相似形状的 Shapelets, 从而增加更多有辨别能力的 Shapelets 用于数据转换, 大大提高了分类的准确度。分别采用 LSSVM、GA-LSSVM 和 AGA-LSSVM 建立分类模型, 共形成 9 种不同的分类方法。选取分类准确率和运行时间两个指标, 对 9 种分类方法的性能进行实验测试对比。实验中采用 8 折交叉验证法将数据集划分为训练集和测试集, 通过训练集对每个分类算法进行迭代训练, 然后在测试集上进行

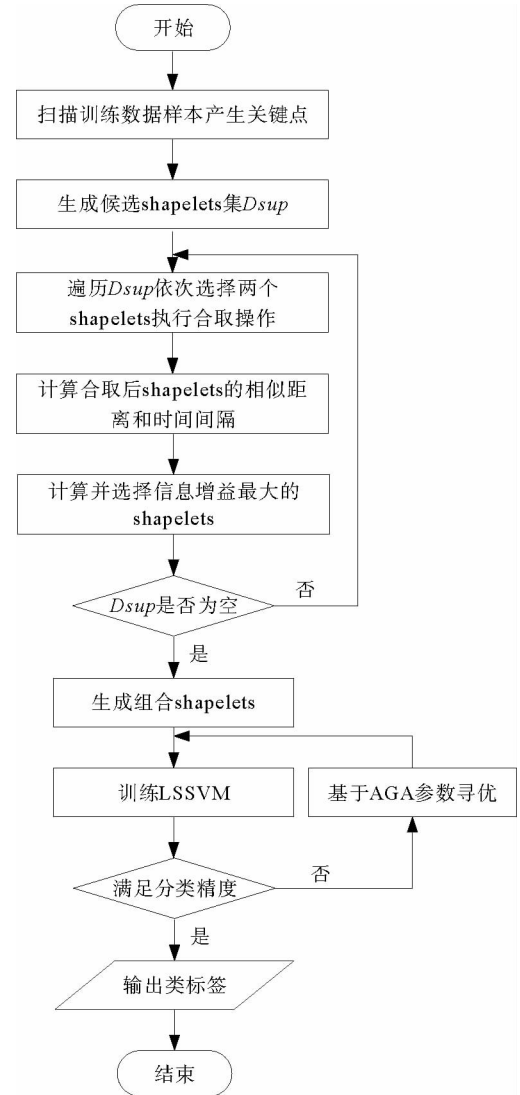


图 1 CSAGA-LSSVM 分类挖掘算法流程图

Fig. 1 Flow chart of CSAGA-LSSVM classification mining algorithm

检验,取平均值作为两个评价指标的最终结果。

1) 分类准确率

分类准确率是测试集上正确识别的样本数与数据对象总数的比值。9 个分类方法在不同数据集上的准确率对比见表 2。

表 2 9 个分类方法在不同数据集上的分类准确率对比

Tab. 2 Comparison of classification accuracy of 9 classification methods on different data sets %

分类方法		数据集				平均准确率
Shapelets 类型	分类模型	Trace	ItalyPower	Wafer	Coffee	
S-Shapelets	LSSVM	85.3	87.4	91.4	88.6	88.2
	GA-LSSVM	88.5	90.3	93.7	89.4	90.5
	AGA-LSSVM	90.8	93.8	96.2	91.7	93.1
L-Shapelets	LSSVM	86.1	88.5	93.6	90.5	89.7
	GA-LSSVM	88.2	91.2	95.2	91.6	91.6
	AGA-LSSVM	91.7	94.6	96.8	92.4	93.9
C-Shapelets	LSSVM	89.5	90.4	93.1	90.7	90.9
	GA-LSSVM	91.6	93.2	95.8	93.6	93.6
	AGA-LSSVM	93.4	96.7	97.5	94.2	95.4

从表 2 可见,CSAGA-LSSVM 分类方法的准确率要优于其他分类方法,且采用 C-Shapelets 的同一分类模型具有最高的准确率,说明 C-Shapelets 通过引入不同 Shapelets 之间的时间间隔对子序列进行组合,并根据信息增益选择最优的子序列,可以有效增强 Shapelets 的辨识能力,提高分类精度。同时基于 C-Shapelets 的 AGA-LSSVM 通过自适应遗传操作,比 LSSVM、GA-LSSVM 能搜索到更优的超参数解,对原始时间序列数据有更完整的特征表示,表现出更好的分类能力。另外,利用 S-Shapelets 的 LSSVM 在 Trace 数据集上的分类准确率较低,说明直接采用单个 Shapelets 对较长的时间序列进行分类,容易忽略不同子序列之间的关联,导致分类效果下降。而 L-Shapelets 采用剪枝策略对相似的 Shapelets 进行过滤,增加其他特征的 Shapelets,具有较高的分类精度。

2) 运行时间

通过调用 Python 中 time 模块的时钟函数返回当前时间,用来计算算法运行的时间。表 3 为 9 个分类方法在不同数据集上的运行时间对比。

表 3 9 个分类方法在不同数据集上的运行时间对比

Tab. 3 Comparison of running time of 9 classification methods on different data sets s

分类方法		数据集			
Shapelets 类型	分类模型	Trace	ItalyPower	Wafer	Coffee
S-Shapelets	LSSVM	21.84	1.42	46.28	18.62
	GA-LSSVM	28.35	4.68	51.74	23.68
	AGA-LSSVM	37.54	4.91	58.16	30.17
L-Shapelets	LSSVM	35.42	3.25	72.84	28.57
	GA-LSSVM	48.14	8.65	81.65	36.84
	AGA-LSSVM	52.94	10.18	90.41	47.15
C-Shapelets	LSSVM	31.45	2.74	51.86	23.17
	GA-LSSVM	36.28	5.86	57.34	29.42
	AGA-LSSVM	43.72	6.37	65.48	35.28

从表 3 可知,直接采用 S-Shapelets 的 LSSVM 分类方法所需时间最少,原因在于 S-Shapelets 没有进行不同 Shapelets 之间的组合运算,而是通过子类划分产生候选 Shapelets,计算过程比 L-Shapelets、C-Shapelets 简单,并且 LSSVM 不进行参数寻优,因此整个算法的运行时间最少。基于 C-Shapelets 的时间序列表示充分利用具有显著特征的关键点,能够快速产生候选 Shapelets,减少了相应子序列的搜索时间,从而算法的运行效率高。L-Shapelets 在所有可能的子序列空间搜索候选 Shapelets,空间复杂度较大,在产生 Shape-

lets 后还要进行迭代剪枝操作,故大多数情况下基于 L-Shapelets 的 AGA-LSSVM 所花费时间较多。另外 AGA-LSSVM 和 GA-LSSVM 采用遗传操作对 LSSVM 超参进行寻优,需要反复迭代直至全局收敛或达到规定的代数,因此运行时间相比传统的 LSSVM 要长,但准确率有所提高。

综上,CSAGA-LSSVM 分类方法在准确率方面具有明显优势,适用于不同规模和密度的数据集,说明所提出的组合 Shapelets 能对原始时间序列进行更完整的特征表示。在运行时间上,基于 C-Shapelets 的 AGA-LSSVM 分类方法比 S-Shapelets 的 LSSVM 长,但比采用 L-Shapelets 的 GA-LSSVM、AGA-LSSVM 运行时间要少。综合考虑分类准确率的影响及分类模型训练的时间要求,CSAGA-LSSVM 分类方法是可行的。

4 基于 CSAGA-LSSVM 的坦克驾驶模拟训练操作数据分析

4.1 数据来源

通过安装在某型坦克驾驶模拟器上的位移传感器、光电传感器实时记录受训人员的操作数据,模拟器系统帧频为 25 fps,每运行 1 帧采集数据 1 次,即每秒采集 25 组数据。采集的操作数据主要包括起动开关、变速装置挡位、加油踏板、离合器踏板等。

为便于分析,对坦克驾驶操作分为基本操作和组合操作。基本操作是指驾驶过程只有一个操纵部件的状态发生变化,是驾驶操作的“元动作”。组合操作是由若干个基本操作按不同的顺序组成的动作序列,与动作次序、完备程度、完成时间等密切相关。换挡操作是适应不同路面环境下的机动要求,是坦克驾驶操作的一项重要内容。下面以“一挡换二挡”为例,运用 CSAGA-LSSVM 算法进行模拟训练结果分析。

在坦克驾驶模拟训练实验中,选择 1 200 名受训人员(二级坦克驾驶员、三级坦克驾驶员、初级坦克驾驶员和无等级人员各 300 名)在某型坦克驾驶模拟教室进行“一挡换二挡”操作训练,每名受训人员进行 6 次换挡操作,通过传感器和模拟训练系统采集到的驾驶员部分操作过程数据如表 4 所示。

表 4 坦克驾驶模拟器“一挡换二挡”操作数据
Tab. 4 Data of the tank driving simulator in “shifting from the first to the second gear” operation

序号	油门踏板 位移/mm	离合器踏板 位移/mm	制动器踏板 位移/mm	左操纵杆 位移/mm	右操纵杆 位移/mm	挡位
1	23.1	3.7	9.5	14.5	28.6	1
2	37.3	4.2	11.2	18.7	32.8	1
3	56.4	4.1	10.8	17.6	24.5	1
4	62.8	4.8	10.5	18.4	15.7	1
5	48.5	5.2	9.8	31.4	14.8	1
...

4.2 最优分类解获取

从表 4 可以看出,油门踏板位移、离合器踏板位移和制动器踏板位移属于一元时间序列数据,变量取值随时间连续变化形成多变量时间序列。本次实验共产生 7 200 个样本,选择油门踏板位移、离合器踏板位移、制动器踏板位移、挡位值为特征变量,对操作成绩离散化处理为类标签 y ,形成决策表,如表 5 所示。

表 5 坦克驾驶模拟器操作数据分析决策表
Tab. 5 Data analysis decision of the tank driving simulator operations

样本	油门踏板位移/mm	离合器踏板位移/mm	制动器踏板位移/mm	挡位	操作成绩
1	43.1 ...	4.7 ...	11.4 ...	1 ...	1
2	54.2 ...	85.8 ...	10.2 ...	1 ...	0
3	32.5 ...	3.8 ...	9.7 ...	2 ...	1
4	23.7 ...	78.4 ...	10.5 ...	1 ...	0
...

从表5中随机选取80%的数据作为训练集进行分类器训练,余下的20%作为测试集。首先通过Shapelets特征提取方法对换挡操作的多维时间序列进行表示,求解组合Shapelets,再依次计算油门踏板位移、离合器踏板位移等时序数据与组合Shapelets之间的距离,实现换挡操作数据变换,用于后续分类输入;然后运用AGA求得最优超参数 $C=65.8$, $\sigma=0.25$ 。基于调优后的分类模型对时序数据进行分类,最终得出分类结果如表6所示。

4.3 结果分析

以表6为依据,结合坦克驾驶换挡操作的理论与实践,对得到的最终解分析如下:

1) “一挡换二挡”操作成绩合格所对应的Shapelets:油门踏板时序数据 $\text{Shapelets}_{\text{acr}}[1] \wedge \text{Shapelets}_{\text{acr}}[2] = \{00,01,00\} \wedge \{00,01\}$,离合器踏板位移时序数据 $\text{Shapelets}_{\text{clu}}[1] \wedge \text{Shapelets}_{\text{clu}}[2] = \{000,010,011,110,111\} \wedge \{111,100,011,010,000\}$,挡位时序数据 $\text{Shapelets}_{\text{gear}} = \{1,2\}$,表示换挡前应平稳踩下油门踏板至适当位置,然后松开油门踏板同时将离合器踏板由初始位置踩到最大位置,迅速将变速杆由1挡推向2挡位置,挂挡后松回离合器踏板应做到前快后稳,同时均匀踩下油门踏板。

2) “一挡换二挡”操作成绩不合格所对应的Shapelets:制动器踏板位移时序数据 $\text{Shapelets}_{\text{brk}} = \{00,01,11\}$,油门踏板时序数据 $\text{Shapelets}_{\text{acr}} = \{00\}$,挡位时序数据 $\text{Shapelets}_{\text{gear}} = \{1\}$,表示“一挡换二挡”过程中踩下了制动器踏板,而未踩油门踏板进行冲车,车辆挡位值没有发生变化;油门踏板时序数据 $\text{Shapelets}_{\text{acr}}[1] \wedge \text{Shapelets}_{\text{acr}}[2] = \{00,01,00\} \wedge \{00,01\}$,离合器踏板位移时序数据 $\text{Shapelets}_{\text{clu}} = \{000\}$,挡位时序数据 $\text{Shapelets}_{\text{gear}} = \{1\}$,表示“一挡换二挡”过程中未踩下离合器踏板,换挡操作不合格。

将“一挡换二挡”操作成绩为合格所对应的Shapelets组合可建立相应的标准操作动作模式,作为“一挡换二挡”操作技能水平的评价等级,对受训人员的驾驶动作进行精确分析。比如,随机选取一名受训人员的训练样本数据进行预处理生成相应的Shapelets,其中离合器踏板位移时序数据 $\text{Shapelets}_{\text{clu}}[1] = \{000,011,100,110\}$,挡位时序数据 $\text{Shapelets}_{\text{gear}} = \{1\}$ 。通过计算与标准操作动作模式之间的距离,可以发现该名受训人员在换挡过程中没有将离合器踏板踩到底,导致挡位状态仍为1挡,挂挡不成功,表明该受训人员掌握动作要领较差,需要在今后的训练中加强理论学习,熟记换挡的关键动作和操作流程,同时注重模拟器基本驾驶动作训练,反复体会换挡要领。

5 结论

本研究为挖掘坦克驾驶模拟训练数据,提出一种CSAGA-LSSVM分类挖掘算法。先根据起始点、结束点、阶跃点、极值点快速获取候选Shapelets,减少产生的候选Shapelets数量;再通过选择距离和时间间隔对Shapelets进行组合操作,并计算信息增益评价最优组合,增强特征的辨识能力,提升分类的准确性;然后设计了自适应遗传算法,通过动态调整交叉和变异概率来加速搜索LSSVM超参数的全局最优解,减少分类挖掘时间;最后,利用CSAGA-LSSVM算法分析某型坦克驾驶模拟器换挡操作数据,提取不同训练水平人员的操作特征,从而更好地发现训练问题,促进个性化训练的开展。

参考文献:

[1] 刘兵,郑怀洲,肖东昀.装备指挥训练模拟数据收集系统研究[J].系统仿真学报,2006,18(增2):246-249.

LIU Bing,ZHENG Huaizhou,XIAO Dongyun.Research on data collection system for equipment command training simulation[J].Journal of System Simulation,2006,18(S2):246-249.

表6 结果数据

Tab. 6 Statistics of Results		
类标签编号	类标签值	对应最优 Shapelets 值
1	合格	$\text{Shapelets}_{\text{acr}} = \{00,01,00\}, \{00,01\}$
		$\text{Shapelets}_{\text{clu}} = \{000,010,011,110,111\}, \{111,100,011,010,000\}$
		$\text{Shapelets}_{\text{gear}} = \{1,2\}$
0	不合格	$\text{Shapelets}_{\text{brk}} = \{00,01,11\}$
		$\text{Shapelets}_{\text{acr}} = \{00\}$
		$\text{Shapelets}_{\text{gear}} = \{1\}$
0	不合格	$\text{Shapelets}_{\text{acr}} = \{00,01,00\}, \{00,01\}$
		$\text{Shapelets}_{\text{clu}} = \{000\}$
		$\text{Shapelets}_{\text{gear}} = \{1\}$
...

- [2]焦楷哲,程培源,刘滔,等.国外军用虚拟训练系统研究[J].飞航导弹,2013(6):60-67.
JIAO Kaizhe, CHENG Peiyuan, LIU Tao, et al. Study on foreign military virtual training systems[J]. Aerodynamic Missile Journal, 2013(6): 60-67.
- [3]ANDERSEN N E, KARL J P, CABLE S J, et al. Vitamin D status in female military personnel during combat training[J]. Journal of the International Society of Sports Nutrition, 2010, 7(1): 38-43.
- [4]ROBINSON M E, TEYHEN D S, WU S S, et al. Mental health symptoms in combat medic training: a longitudinal examination[J]. Military Medicine, 2009(6): 572-576.
- [5]CADY A. Using the National Training Center instrumentation system to aid simulation-based acquisition[D]. Santa Monica: Pardee Rand Graduate School, 2017.
- [6]WANG J, LIN Y, HOU S Y. Data mining approach for training evaluation in simulation-based training[J]. Computers & Industrial Engineering, 2015, 80: 171-180.
- [7]邓桂龙,刘智慧,贾志东.作战仿真实验数据关联规则挖掘研究[J].军事运筹与系统工程,2008,22(4):46-50.
DENG Guilong, LIU Zhihui, JIA Zhidong. Research on data association rules mining in combat simulation experiment[J]. Military Operations Research and Systems Engineering, 2008, 22(4): 46-50.
- [8]唐志武,薛青,刘翔.基于决策树的装甲兵登陆突击仿真试验数据挖掘[J].装甲兵工程学院学报,2008,22(1):6-9.
TANG Zhiwu, XUE Qing, LIU Xiang. Data mining in the simulation tests of armored forces' landing attack based on the decision tree[J]. Journal of Academy of Armored Force Engineering, 2008, 22(1): 6-9.
- [9]柴慧敏,赵昀瑶,方敏.利用先验正态分布的贝叶斯网络参数学习[J].系统工程与电子技术,2018,40(10):2370-2375.
CHAI Huimin, ZHAO Yunyao, FANG Min. Learning Bayesian networks parameters by prior knowledge of normal distribution[J]. Systems Engineering and Electronics, 2018, 40(10): 2370-2375.
- [10]杨瑞,王萍,索瑞霞,等.数据挖掘技术在 CRM 中的应用研究—基于决策树算法[J].中国管理信息化,2019,22(15):53-55.
YANG Rui, WANG Ping, SUO Ruixia, et al. Application of data mining technology in CRM based on decision tree algorithm[J]. China Management Information, 2019, 22(15): 53-55.
- [11]吴雨茜,王俊丽,杨丽,等.代价敏感深度学习研究方法研究综述[J].计算机科学,2019,46(5):1-12.
WU Yuxi, WANG Junli, YANG Li, et al. Survey on cost-sensitive deep learning methods[J]. Computer Science, 2019, 46(5): 1-12.
- [12]业巧林,许等平,张冬.基于深度学习特征和支持向量机的遥感图像分类[J].林业工程学报,2019,4(2):119-125.
YE Qiaolin, XU Dengping, ZHANG Dong. Remote sensing image classification based on deep learning features and support vector machine[J]. Journal of Forestry Engineering, 2019, 4(2): 119-125.
- [13]LANG H, ARNOLLD M. Numerical aspects in the dynamic simulation of geometrically exact rods[J]. Applied Numerical Mathematics, 2012, 62(10): 1411-1427.
- [14]孙英娟.基于粗糙集的分类方法研究[D].长春:吉林大学,2011.
SUN Yingjuan. Research on classification methods based on Rough Set[D]. Changchun: Jilin University, 2011.
- [15]LINES J, DAVIS L M, HILLS J, et al. A shapelet transform for time series classification[C]//18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, 2012: 289-297.
- [16]WANG H, LI C, SUN H, et al. Shapelet classification algorithm based on efficient subsequence matching[J]. Data Science Journal, 2018, 17(5): 8-20.
- [17]ARIAS M, TRONCOSO A, RIQUELME J C. A kernel for time series classification: application to atmospheric pollutants[J]. Advances in Intelligent Systems and Computing, 2013, 188: 417-426.
- [18]嵇存.基于 Shapelet 的时间序列分类方法研究[D].青岛:山东大学,2017.
JI Cun. Time series classification methods based on shapelet[D]. Qingdao: Shandong University, 2017.
- [19]NORBERT K, CHRISTOPHER G, JUSTUS P, et al. Object-action complexes: grounded abstractions of sensory-motor processes[J]. Robotics and Autonomous Systems, 2011, 59(10): 740-757.