doi:10.11835/j.issn.1000-582X.2023.07.007

一种基于数据驱动的动态时序分类算法

赵庶旭,张家祯,王小龙,张占平
(兰州交通大学 电子与信息工程学院,兰州 730070)

摘要:针对物联网时序数据中存在的数据冗余现象和动态信息难以捕捉的问题,提出了一种基于数据驱动的动态时序分类算法。通过动态内部主元分析法(dynamic internal principal component analysis,DiPCA)提取传感设备采集的时间序列中的动态信息,实现降维及提炼动态信息的作用;利用麻雀搜索算法优化分类算法参数,强化支持向量机(support vector machines,SVM)算法性能并使其对含有shapelet局部特征的时序特征进行建模,最终构成双向演进算法框架,实现时序分类功能。利用UCR时序数据集和边缘计算模拟数据检验该算法的性能,结果表明,与基本算法相比,该算法的综合性能明显提高,并验证算法分类功能在仿真环境中的有效性与优越性。

关键词:数据驱动;动态内部主元分析法;shapelet;麻雀搜索算法;支持向量机;时间序列分类
中图分类号:U448.213 **文献标志码:**A **文章编号:**1000-582X(2023)07-063-12

A data-driven dynamic time series classification algorithm

ZHAO Shuxu, *ZHANG Jiazhen*, *WANG Xiaolong*, *ZHANG Zhanping*
(School of Electronic and Information Engineering, Lanzhou Jiaotong University,
Lanzhou 730070, P. R. China)

Abstract: Aiming at the problems of data redundancy and difficulty in capturing dynamic information in IoT time series data, this paper proposes a data-driven dynamic time series classification algorithm. The dynamic information in the time series collected by sensing devices is extracted by DiPCA (dynamic internal principal component analysis) to realize the role of dimensionality reduction and refining dynamic information; the parameters of the classification algorithm are optimized by using the sparrow search algorithm to enhance the performance of the SVM algorithm and make it model the temporal features containing shapelet local features, which finally constitutes a two-way evolutionary algorithm framework to realize the temporal classification function. The performance of the algorithm is examined using UCR temporal data set and edge computing simulation data, and the results show that the comprehensive performance of the algorithm is significantly improved compared with the basic algorithm, and the effectiveness and superiority of the classification function of the algorithm in the simulation environment is verified.

Keywords: data-driven; dynamic internal principal component analysis method; shapelet; sparrow search algorithm; support vector machine; time series classification

收稿日期:2021-09-29
基金项目:甘肃省重点研发计划项目(20YF8GA123)。
Supported by the Key Research and Development Program of Gansu Province (20YF8GA123).
作者简介:赵庶旭(1976—),男,教授,博士,主要从事智能交通、边缘计算、深度学习、目标检测方向研究,(E-mail) zhaosx@mail.lzjtu.cn。
通信作者:张家祯(1997—),男,硕士生,主要从事数据分析、边缘计算、深度学习、异常检测方向研究,(E-mail) 1603151510@qq.com。

近年来,物联网技术飞速发展,如何为物联网提供高效的时序数据挖掘方案已成为研究热点。充分地将数据中的重要信息提取、整理,才能规划出准确,高效的模型。而应用型数据规模正持续地以指数级规模上升,数据的属性也在不断的扩张。因而对于数据集,通常采取降低维度的方式,简化数据规模,主成分分析法(principal component analysis, PCA)通过研究属性之间的线性关系,筛选出独立性极强的属性集,去除了多余的冗余属性,降低了分析数据的难度,保留了数据的重要信息。基于此原理,Ku等[1]提出了动态主成分分析法(dynamic principal component analysis, DPCA),该方法将原始属性通过时间窗口迭代读取,形成增广数据矩阵,利用PCA进行过程监测,以此达到折叠数据并以时序角度观察信息;该方法扩张了数据规模,易造成维度灾难,且难以挖掘数据相互关联程度。Li等[2]提出了动态潜变量算法,首次提出自回归PCA算法,要求最大化自协方差的同时提取潜变量,并以此建立自回归模型;该方法提取潜变量一阶次的自协方差,难以挖掘更深层次的动态关系。Dong等[3] 提出动态内部主元分析算法(dynamic internal principal component analysis, DiPCA),认为优化方向在于通过动态关联状态作为分析重心,在已有的动态性数据中,量化数据的自相关性,寻找多维主元内部的相关联系,建立目标函数,最大化已提取的动态主元与预测的一般时刻动态主元的协方差,最终达到提取动态主元的目标。

2009年Keogh提出shapelet象形子序列,采取信息增益规则提取子序列代表时序的局部关键特征,相比于最近邻算法(nearest neighbor algorithm,NNA) ,以shapelet构造决策树的分类方法准确率有所提高,但面对多分类问题仍有不足之处。原继东等[4]提出了基于shapelet的剪枝和覆盖算法,减少不必要的数据特征;闫欣鸣等[5]提出了趋势特征符号化方法表示时序的趋势信息,提升运算效率,但是数据分类效果并不稳定。综上所述,可考虑添加潜在变量信息丰富子序列特征内容,产生更优质的辨识性分类属性。Coloni等[6]提出的蚁群优化(ant colony optimization,ACO)以信息素的高低指导蚁群寻找最优路径,鲁棒性高,适合求解连续性函数,但因难以统一多样性信息,极易陷入局部最优状态。Karaboga等[7]以蜜蜂采蜜的工作习性作为参考,不同分工的工蜂群以正负2种信号反馈获取最优解位置,但信号传播较慢,导致算法不易收敛。Xue等[8]提出的麻雀搜索算法在蜂群算法,基础上加入新的探测机制,搜索食物密集处的同时,关注安全程度,筛选满意解集合,获取全局最优解。

SVM在研究动态数据分类有显著作用,具有较好的鲁棒性和有效规避维数灾难等优点,得到了学者重视。多变量问题的解决思想最为直观的便是统一线性化,但使用简易的线性化模型有可能出现过拟合现象,或者是过度简化变量相关性质。文献[9]提出基于最小二乘支持向量机(least squares support vector machine,LS-SVM)的相关性局部即时模型优化,以解决模型误差大、训练学习精度较低等问题。文献[10]运用遗传算法(genetic algorithm,GA)优化LS-SVM模型关键参数,提高模型的收敛能力,但未深度挖掘数据的时序规律特征。

文献[11-13]从分解流程、提取惰性信息、时序大数据挖掘3个角度分别研究提取大数据的核心时序信息的方法,有效压缩了信息容量,但都难以提取典型动态特征支撑分类算法。文献[14]使用shapelet快速完成分类运算,文献[15-16]引入了卷积算法有效提取更精准的信息特征,但是忽略了静态冗余信息可能会影响分类的结果。提出的DiPCA-Shapelet-SSA-SVM组合模型重点研究数据信息的内联性与外部相关扩展性,按照固定时间单位对时序数据进行划分,每单位时间内拥有相同的探测数据数量,且探测时间一致;因此,设置探测值作为属性向量,转化单维数据为多维数据,提取有效潜变量,并获取综合动态指标与静态检测指标,提取的动态指标作为新的参考序列供shapelet子序列提取,获得典型时序特征;受自然界生物世界的启发,群体智能优化算法模拟生物行为,作为搜索算法的运行机制,达到优化解空间全局的效果;麻雀搜索算法优化向量机核心因子,充分运用动态特征。实验结果表明本文算法性能指标优于现有常用分类算法,准确率与召回率明显提升,并在仿真实验中得到有效验证。

1　原理介绍

1.1　特征提取方法

1.1.1　DiPCA

PCA难以满足时序信息的预处理工作,时间序列与时间联系密切,其动态特性无法只靠特征贡献率降维和清理数据的方式获取,DiPCA引入自回归模型,设立协方差目标函数,在迭代过程中提取有效动态变量,重构新的时序信息,完整保留原有的信息内容。

设置$\boldsymbol{x}_k$为t时刻采集获取的时序数据,$\boldsymbol{w}$作为均衡因子向量,建立动态主元$\boldsymbol{r}_t$,

$$\boldsymbol{r}_t=\boldsymbol{x}_t\boldsymbol{w}。\tag{1}$$

引用自回归模型,划分已建立的动态主元作为训练集训练自回归模型系数$\boldsymbol{m}_i$,提取有效动态特征,得模型

$$\boldsymbol{r}_t=\boldsymbol{m}_1\boldsymbol{r}_{t-1}+\boldsymbol{m}_2\boldsymbol{r}_{t-2}+\cdots+\boldsymbol{m}_s\boldsymbol{r}_{t-s}+\boldsymbol{e}_t,\tag{2}$$

式中:$\boldsymbol{m}$和$\boldsymbol{w}$为单位向量,即$\|\boldsymbol{m}\|=1,\|\boldsymbol{w}\|=1$,$\boldsymbol{e}_t$为噪声向量,$s$为序列阶次量。

由此得动态主元预估模型

$$\hat{\boldsymbol{r}}_t=[\boldsymbol{r}_{t-1}{}^{\mathrm{T}}\boldsymbol{r}_{t-2}{}^{\mathrm{T}}\cdots\boldsymbol{r}_{t-s}{}^{\mathrm{T}}](\boldsymbol{m}\otimes\boldsymbol{w})。\tag{3}$$

设置向量组$\boldsymbol{X}_i=\left[\boldsymbol{x}_i\boldsymbol{x}_{i+1}\cdots\boldsymbol{x}_{N+i-1}\right]^{\mathrm{T}}$,其中$i=1,2,\cdots,s+1$,$N$为序列理想切割划分单位,通过$\boldsymbol{X}_i$递推式建立新的阶次向量$\boldsymbol{Z}_s=\left[\boldsymbol{X}_1\boldsymbol{X}_2\cdots\boldsymbol{X}_s\right]$。

设立目标函数,将预估动态主元与原动态主元两者间的协方差达到最大化,

$$\max_{\boldsymbol{w},\boldsymbol{m}}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}_{s+1}^{\mathrm{T}}\boldsymbol{Z}_s(\boldsymbol{m}\otimes\boldsymbol{w}),\tag{4}$$

$$\text{s.t.}\|\boldsymbol{w}\|=1,\|\boldsymbol{m}\|=1。\tag{5}$$

采用拉格朗日算子法,确定参数$\boldsymbol{w}$、$\boldsymbol{m}$与$\boldsymbol{r}_i$之间关联方程如下:

$$\lambda_{\boldsymbol{w}}\boldsymbol{w}=\sum_{i=1}^{s}\boldsymbol{m}_i\left(\boldsymbol{X}_{s+1}^{\mathrm{T}}\boldsymbol{r}_i+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{r}_{s+1}\right),\tag{6}$$

$$\lambda_{\boldsymbol{m}}\boldsymbol{m}=\left[\boldsymbol{r}_1\boldsymbol{r}_2\cdots\boldsymbol{r}_s\right]^{\mathrm{T}}\boldsymbol{r}_{s+1}。\tag{7}$$

通过迭代循环模型,设置收敛系数,循环计算均衡因子以及自回归权重系数,直至满足收敛要求。

1.1.2　Shapelet子序列介绍

有效区分不同类别的序列的算法需要选取最佳子序列,shapelet搜索算法实质是计算子序列与母序列间的最短距离,以此判断所选序列是否代表典型局部特征。采用欧式距离计算如下:

$$\text{Dist}(\boldsymbol{T},\boldsymbol{Q})=\sqrt{\sum_{i=1}^{v}(t_i-q_i)^2},\tag{8}$$

$$\text{SubseqDist}(\boldsymbol{U},\boldsymbol{T})=\min(\text{Dist}(\boldsymbol{U},\boldsymbol{T}_i))。\tag{9}$$

式中:$\boldsymbol{T}$、$\boldsymbol{Q}$为2条时间序列;t_i、q_i分别为前后时间序列的单位时间内有序数列数值。通过公式(9)可确定子序列$\boldsymbol{U}$与母序列$\boldsymbol{T}$的距离,两者距离数值作为判断信息熵的重要指标,直接决定了shapelet的匹配程度。

1.2　SVM模型参数优化

1.2.1　麻雀搜索算法(SSA)

受麻雀生活习性启发,设置$\boldsymbol{X}_{i,j}^{t+1}$为麻雀位置向量,建立3类种族行动群体,分别是发现者、追随者、侦察者。以下分别介绍其运行原理及特点。

1)发现者模式

$$\boldsymbol{X}_{i,j}^{t+1}=\begin{cases}\boldsymbol{X}_{i,j}\exp\left(-\dfrac{i}{\alpha\cdot\text{iterations}-\max}\right), & \text{if } R_2<\text{ST};\\ \boldsymbol{X}_{i,j}+\boldsymbol{Q}\times\boldsymbol{L}, & \text{if } R_2\geqslant\text{ST};\\ i\leqslant I,j\leqslant J。\end{cases}\tag{10}$$

该公式以规划发现者群体在第t代中第i个单体的第j维位置,根据正态分布矩阵$\boldsymbol{Q}$与预定距离$\boldsymbol{L}$的乘积

量确定侦察者的大致分布，设ST为警戒阈值，以随机数值R_2与ST的对比关系确定单体的行动策略。

2)跟随者模式

$$\boldsymbol{X}_{i,j}^{t+1}=\begin{cases}\boldsymbol{Q}\cdot\exp\left(\dfrac{\boldsymbol{X}_{\text{worst}}-\boldsymbol{X}_{i,j}^{t}}{i^2}\right), & \text{if } i>n/2;\\ \boldsymbol{X}_{P}^{t+1}+\left|\boldsymbol{X}_{i,j}-\boldsymbol{X}_{P}^{t+1}\right|\times\boldsymbol{A}^{+}+\boldsymbol{L}, & \text{otherwise}。\end{cases}\tag{11}$$

当i小于麻雀总数量的1/2，说明当前的麻雀单位处于缺乏能量的饥饿状态，因此测定当前最劣单位$\boldsymbol{X}_{\text{worst}}$与自身单位的距离，并将其投射到指数函数中，再根据随机分布矩阵$\boldsymbol{A}^{+}$分配随机步长向周围搜寻捕食点。

3)侦察者模式

$$\boldsymbol{X}_{i,j}^{t+1}=\begin{cases}\boldsymbol{X}_{\text{best}}^{t}+\beta\left|\boldsymbol{X}_{i,j}^{t}-\boldsymbol{X}_{\text{best}}^{t}\right|, & \text{if } f_{\text{i}}>f_{\text{g}};\\ \boldsymbol{X}_{i,j}^{t}+K\cdot\left(\dfrac{\left|\boldsymbol{X}_{i,j}^{t}-\boldsymbol{X}_{\text{worst}}^{t}\right|}{(f_{\text{i}}-f_{\text{r}})+\varepsilon}\right), & \text{if } f_{\text{i}}=f_{\text{g}}。\end{cases}\tag{12}$$

式中：β为符合标准正态分布的步长参数；K为[−1,1]的满足均匀分布的随机数。以麻雀单位的适应值f_i、全局最优麻雀的适应值f_{g}、目前最差位置麻雀的适应值f_{r}之间的关系直接确定侦察者的移动策略。侦察的麻雀已处于当前的最优位置$\boldsymbol{X}_{\text{best}}^{t}$时，该单位$\boldsymbol{X}_{i,j}^{t}$会逃离到自身附近的一个位置；反之，该麻雀并未处于最优位置，它会逃到当前最优位置的邻域范围内。

以上3类群体互相牵制，不断调整、更新最优解位置。与其他群智能算法相比，SSA在集中搜寻满意解的同时，引入侦察机制，避免陷入局部最优解问题。

1.2.1 SSA-SVM算法

SVM分类数据集时以确定超平面分割数据集。当数据集容量较大或者种类较多时，SVM需要迭代划分出新的超平面处理信息，准确度会有所下降。SSA算法可通过调整罚函数因子以及核函数参数有效提高该算法的分类效果。SVM算法如式(13)~(15)所示。

$$\min\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{v}\boldsymbol{\zeta}_i\ ;\tag{13}$$

$$\begin{cases}\boldsymbol{y}^{(i)}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)}+\boldsymbol{b}\right)\geqslant 1-\boldsymbol{\zeta}_i,\\ \boldsymbol{\zeta}_i\geqslant 0;\end{cases}\tag{14}$$

$$K\left(x,y\right)=e^{-\gamma\|x-y\|^2}。\tag{15}$$

式中，C为罚函数因子，调控惩罚向量$\boldsymbol{\zeta}$对权重$\boldsymbol{w}$的松弛化操作程度。公式(14)为$\boldsymbol{\zeta}$对目标函数方程$\boldsymbol{y}=\boldsymbol{wx}+\boldsymbol{b}$的松弛化限制原理介绍。公式(15)为高斯核函数公式，用于扩增数据的属性维度[17]，在新生成的属性空间构造超平面，解决多分类数据难分解的问题[18]，参数γ为高斯核函数因子，决定了维度上升后坐标单位的划分长度。

2 算法框架介绍

2.1 时序数据预处理

时序数据按组元为单位，以时间作为有序规则，每组数据时序集合同等划分为相同数量属性指标，一维传感信号依照时间进程转化为不同时刻的状态数据，再以DiPCA算法剔除多余的静态数据集合，抽取带有动态特征的集合(图1)，固定探测信号的准确时刻，不同元组同一时刻采集数据，作为自身属性衡量分数，再以降维工具对信息集采用一种近似于数据蒸馏的方式，蒸馏静态信息，以此达到简化数据量并保留有效信息的目的。该算法在迭代过程中，通过自回归方程不断提取动态特征，方程的阶次值由数据交叉训练中确定；数据不断抛去的静态数据集在迭代收敛结束后作为重要检验指标判断生成的动态主元能否满足预估期望。如果剩余数据可达到传统PCA算法主元贡献率的理论阈值，则可推算已满足算法要求。

其中TSD(time series data)为原始时序集，预设置最大主元数目num_d以及自回归方程阶次数s，提取部分原始数据集分批训练模型(如图1所示)，每一次迭代过程中获取新的主元$\boldsymbol{r}_i$，直至迭代结束或满足期望主

元数目。$\boldsymbol{T}_{\text{new}}$为已处理的特征时序集合。

表 1　DiPCA 特征提取算法流程

Table 1　The algorithm flow of DiPCA features extraction

算法 1	DiPCA 特征提取算法
输入:	TSD
中间变量:	$\boldsymbol{m}$、$\boldsymbol{w}$、$\boldsymbol{r}$、$\boldsymbol{p}$、$\boldsymbol{T}$、iter
输出:	$\boldsymbol{T}_{\text{NEW}}$
	Begin
	$\boldsymbol{T}$ = zero − meannormalization(TSD):$\boldsymbol{w}$ = random[0,1]
	While i<num_d
	While iter > ε;
	$\boldsymbol{r} = \boldsymbol{T}\boldsymbol{w}_i$
	$\boldsymbol{m} = \left[\boldsymbol{r}_1\ \boldsymbol{r}_2 \ldots \boldsymbol{r}_s\right]^{\mathrm{T}} \boldsymbol{r}_{s+1}$
	$\boldsymbol{w}_{\text{new\_i}} = \sum_{i=1}^{S} \boldsymbol{m}_i \left(\boldsymbol{T}_{S+1}{}^{\mathrm{T}} \boldsymbol{r}_i + \boldsymbol{T}_i{}^{\mathrm{T}} \boldsymbol{r}_{s+1}\right)$
	$\boldsymbol{w}_{\text{new\_i}} = \boldsymbol{w}_{\text{new\_i}} / \left\| \boldsymbol{w}_{\text{new\_i}} \right\|$;
	$\boldsymbol{m} = \boldsymbol{m} / \left\| \boldsymbol{m} \right\|$; iter = $\left\| \boldsymbol{r} - \boldsymbol{T}\boldsymbol{w}_{\text{new\_i}} \right\|$;
	$\boldsymbol{T} = \boldsymbol{T} - \boldsymbol{r}\boldsymbol{p}_i{}^{\mathrm{T}}$; $\boldsymbol{p}_i = \boldsymbol{T}{}^{\mathrm{T}}\boldsymbol{r} / \boldsymbol{r}^{\mathrm{T}}\boldsymbol{r}$;
	end;
	end;
	End;
	$\boldsymbol{T}_{\text{NEW}} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \cdots, \boldsymbol{r}_i\}$;

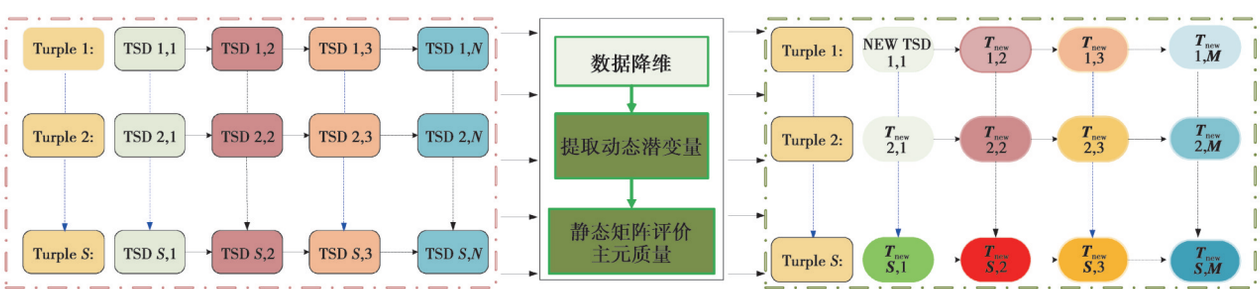

图 1　DiPCA 数据转化示意图

Fig. 1　DiPCA data conversion

在数据转化的过程中,元组 Turple {1,2,...S}每单位对应一条时序变量,变量经过统一切割,生成以时间顺序作为划分依据的多维属性,再经过该算法迭代获取动态主元属性。元组会匹配到对应自身的多元变量,其本质仍为时序值,带有明显的时间流动规律。

2.2　Shapelet 子序列提取

已获取的新型时间序列采用 shapelet 搜索算法为单位时间序列提供可靠子序列集合,在该算法中采用欧式距离法衡量序列间的最短距离。计算子序列中可以代表母序列的显著局部特征的 shapelet 序列[19]。

首先限制 shapelet 的长度范围,经验上最大长度为数据集中最小长度的序列对应的长度数值。滑动窗口在设定的长度范围内截取数据集中所有的序列 S,等效于子序列全集。以信息增益(评价指标为互信息素)作为评判子序列的质量标准,具有最高信息增益值的序列即为 $\boldsymbol{T}_{\text{NEW}}$ shapelets,而信息增益最大的那一个便是 shapelet。

表2 shapelet子序列转换算法流程

Table 2 The algorithm flow of shapelet subsequence transformation

算法 2	shapelet子序列转换算法
	$\boldsymbol{T}_{\mathrm{NEW}}$;
输入：	min=min(shapelet_length);
	max=max(shapelet_length);
中间变量：	$\boldsymbol{S}$、s、DS、quality、 $\boldsymbol{T}_{\mathrm{NEW}}$ shapelets
输出：	$\boldsymbol{S}_{\mathrm{NEW}}$;
	quality (pre - gain threshold); $\boldsymbol{T}_{\mathrm{NEW}}$ shapelets ← ∅,
	for all $\boldsymbol{T}_{\mathrm{new\_i}}$ in $\boldsymbol{T}_{\mathrm{new}}$ do
	$\boldsymbol{S}$←shapelet candidates ($\boldsymbol{T}_{\mathrm{new\_i}}$, min, max);
	for all subsequences s in $\boldsymbol{S}$ do
	DS←calculate distances(s, $\boldsymbol{T}_{\mathrm{new\_i}}$)
	quality←evaluate shapelets ($\boldsymbol{S}$, DS)
	end for
	$\boldsymbol{T}_{\mathrm{NEW}}$ shapelets.add ($\boldsymbol{S}$, quality)
	end for
	$\boldsymbol{S}_{\mathrm{NEW}}$←∅
	for all $\boldsymbol{T}_{\mathrm{new\_i}}$ in $\boldsymbol{T}_{\mathrm{NEW}}$ do
	for all $\boldsymbol{S}_j$ in $\boldsymbol{T}_{\mathrm{NEW}}$ shapelets do
	$\boldsymbol{S}_{\mathrm{NEW}}ij$ = dist ($\boldsymbol{T}_{\mathrm{new\_i}}$, $\boldsymbol{S}_j$)
	end for
	end for
	return $\boldsymbol{S}_{\mathrm{NEW}}$ where $\boldsymbol{S}_{\mathrm{NEW}}$ is a 2D matrix

为支持SSA-SVM模型的训练学习，计算shapelet集合以与原序列的欧式距离作为投放分类训练器的时序属性列表。输出的$\boldsymbol{S}_{\mathrm{NEW}}$为类别数量为$k$,长度为$n$的时序属性数据。

2.3 数据驱动的SSA-SVM分类一体化流程图

SSA-SVM模型以具有时间序列特征的数据作为多维属性集划分超平面边界，SSA寻找最适合当前数据集的参数因子、惩罚函数因子和核函数因子来构造状态矩阵，优化目标是最大化由精确率和召回率的调和平均值构成的F_1分数，该分数由准确率和召回率组成，同时考虑全局操作精度和计算成本，并在周期内更新参数，调整SVM算法的边界范围。

如上所述，在麻雀搜索算法的侦察模型中，由于步长参数不能随搜索范围灵活调整，而框架是基于数据驱动的理论优化条件，需要精确定位可行解的邻域分布，所以调用惯性权重来更新该参数的选择范围[20-21]。随着迭代次数的积累，步长参数会降低选择范围，可获得良好的局部搜索能力，弥补了侦察者参数选择的不稳定性，有目的地使侦察者行动路径所传达的搜索范围更接近自身的目标函数要求(图2)。

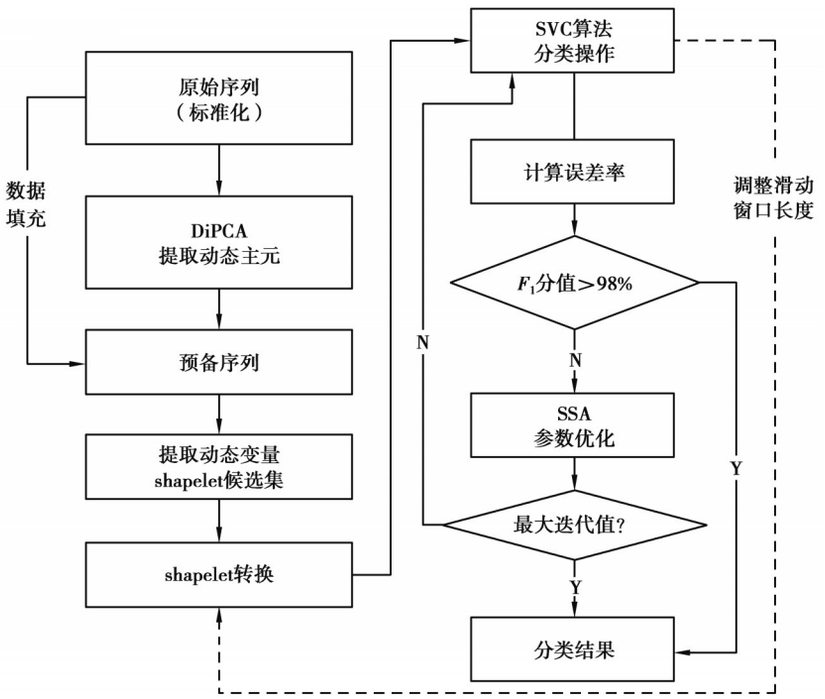

图 2　组合模型优化 SSA-SVM 分类一体化流程图

Fig. 2　Combination model optimization SSA-SVM classification integration framework diagram

3　仿真与分析

3.1　实验说明

选取 10 类 UCR 时间序列分类数据集验证算法效果,抽取显著异或特征,并与现有 6 种分类算法对比,验证新型组合算法的有效性和提升效果;选取中国西北部某城市 2018~2020 年 $PM_{2.5}$ 空气质量序列数据,实验前已将该数据标签化,以空气质量指数划分为优、良、中、差 4 类空气状况。设计边缘计算服务器分配布局,模拟序列数据上传,检测边缘服务器数据分类精确度,并衡量整体计算框架的成本与消耗。

3.2　UCR 时间序列数据集实验对比

根据 UCR 数据分类集,选取能代表序列类别的序列特征,图 3~8 分别为数据集 Gun_Point、Coffee、MoteStrain 原始数据图像以及经算法获得的特征序列。

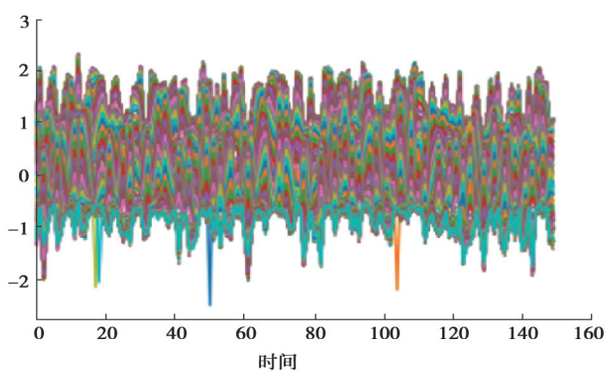

图 3　Gun_Point 原始序列

Fig. 3　Gun_Point original time series

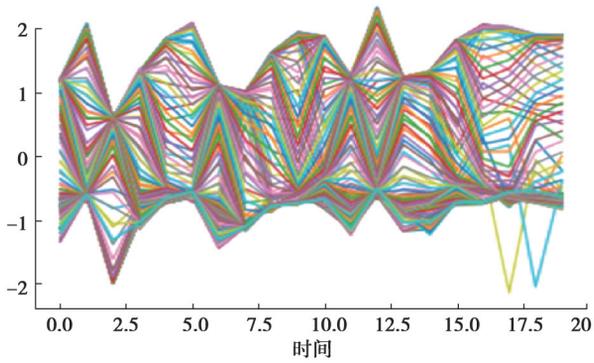

图4 Gun_Point子序列

Fig. 4 Gun_Point time subsequence

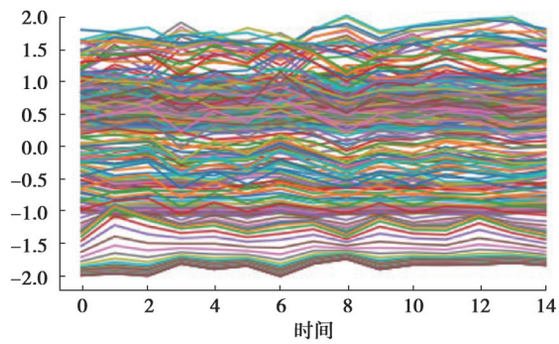

图5 Coffee原始序列

Fig. 5 Coffee original time series

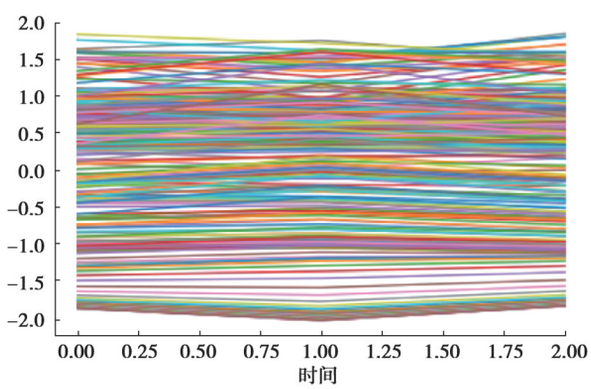

图6 Coffee子序列

Fig. 6 Coffee time subsequence

以数据集Gun_Point、Coffee举例说明，数据集容量较大时难以去除冗余信息，同一时间单位内存在多余的序列信息，复杂的交叉信号点导致难以收集不同种类的典型特征。shapelet可有效抽取局部信息，避免相似数据出现重叠现象。数据集MoteStrain部分数据明显异与整体序列规律趋势，无规律异常值不具有代表性，故采用DiPCA可以在完成降维工作的同时，通过统计学指标去除异常数据，达到清洗数据、蒸馏信息的目的。

表3展现的是算法表现最好的6种数据集的F_1分数对比结果，其中对比算法包括支持向量机(support vector machine，SVM)、最近邻算法1NN(1 nearest neighbor)、朴素贝叶斯(naive Bayes，NB)、决策树C4.5，以及改进型算法加权随机森林(weighted random forest，WRF)与新型深度学习算法CNN-LSTM。通过表3数据可以看出，改进的算法与SVM相比，准确率与召回率整体有明确上升。Coffee数据集容量较大，在这里提取

动态变量的效果显著，提取出重要序列进行比较，F_1值提到了 1.37%；数据集 FaceFour 属于长序列数据，shapelet 收集数据间存在的局部差异，分类器集中处理异或特征，F_1值提高了 2.9%。

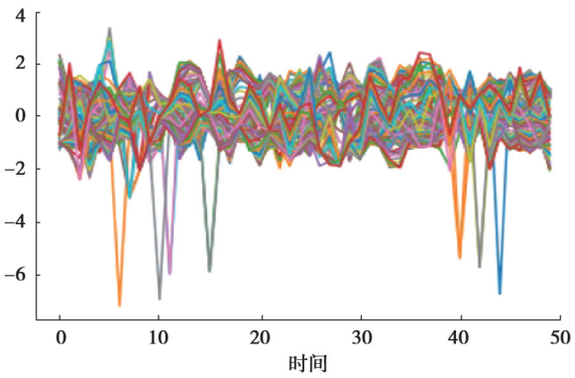

图 7　MoteStrain 原始序列

Fig. 7　MoteStrain original time series

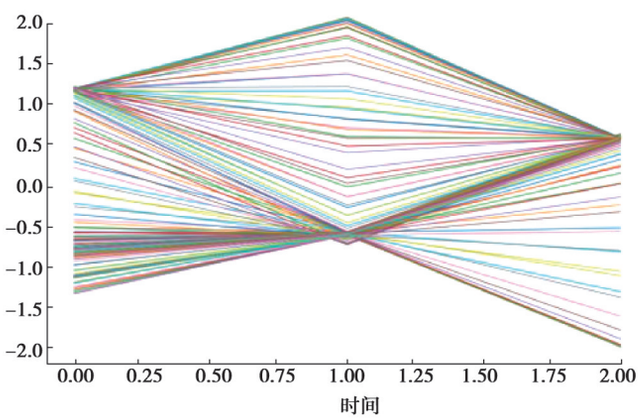

图 8　MoteStrain 子序列

Fig. 8　MoteStrain time subsequence

表 3　不同模型的 F_1 分数

Table 3　F_1 scores of different models

数据集	本文算法	SVM	1NN	NB	WRF	C4.5	CNN-LSTM
Gun_Point	**86.45**	80.21	91.21	78.20	94.39	76.98	86.67
Coffee	**98.20**	96.83	75.00	67.76	85.33	56.89	93.01
MoteStrain	**89.24**	86.32	85.72	85.20	86.90	78.64	84.21
FaceFour	**90.23**	87.33	87.52	84.21	78.12	71.62	80.37
CBF	**91.03**	88.34	85.13	88.24	89.11	67.42	88.51
Trace	**1.00**	1.00	82.34	97.22	95.30	95.32	96.20

3.3　$PM_{2.5}$ 空气质量分类

设计仿真实验场景，设置 100 个距基站 200 m 范围内随机均匀分布传感器节点，50 个边缘服务器均匀分布在 8 个相同面积的区域内，假设已知带宽 35 MHz，，传感器的数据上传速度为 6.8 Mbps，上传功率为 1 800 MW，设备空闲状态时放电功率为 200 MW，边缘传感设备的极限 CPU 主频为 2 GHz。分配中国西北部某城市 2018-2020 年 $PM_{2.5}$ 数据作为传感节点获取的时序集合，采用直接传输方式与中继转发方式将数据传输至边缘服务器，经计算获取分类结果，传输至服务云端。

收集的数据通过 DiPCA-Shapelet 处理，提取的特征要求明显异于其他种类数据，相同性质的时序元素无法起到分类作用，因此，需要去除无显著作用的属性值。通过 DiPCA 算法结合文献[3]使用的 3 种统计指标

(Q统计量,T统计量以及霍林斯指标)确定$PM_{2.5}$异常情况高发时间段,根据真实空气质量状况筛选出误差型异常点集合,并确定保留异或性明显的特征因子,处理结果的可视化图像如图9和图10所示。

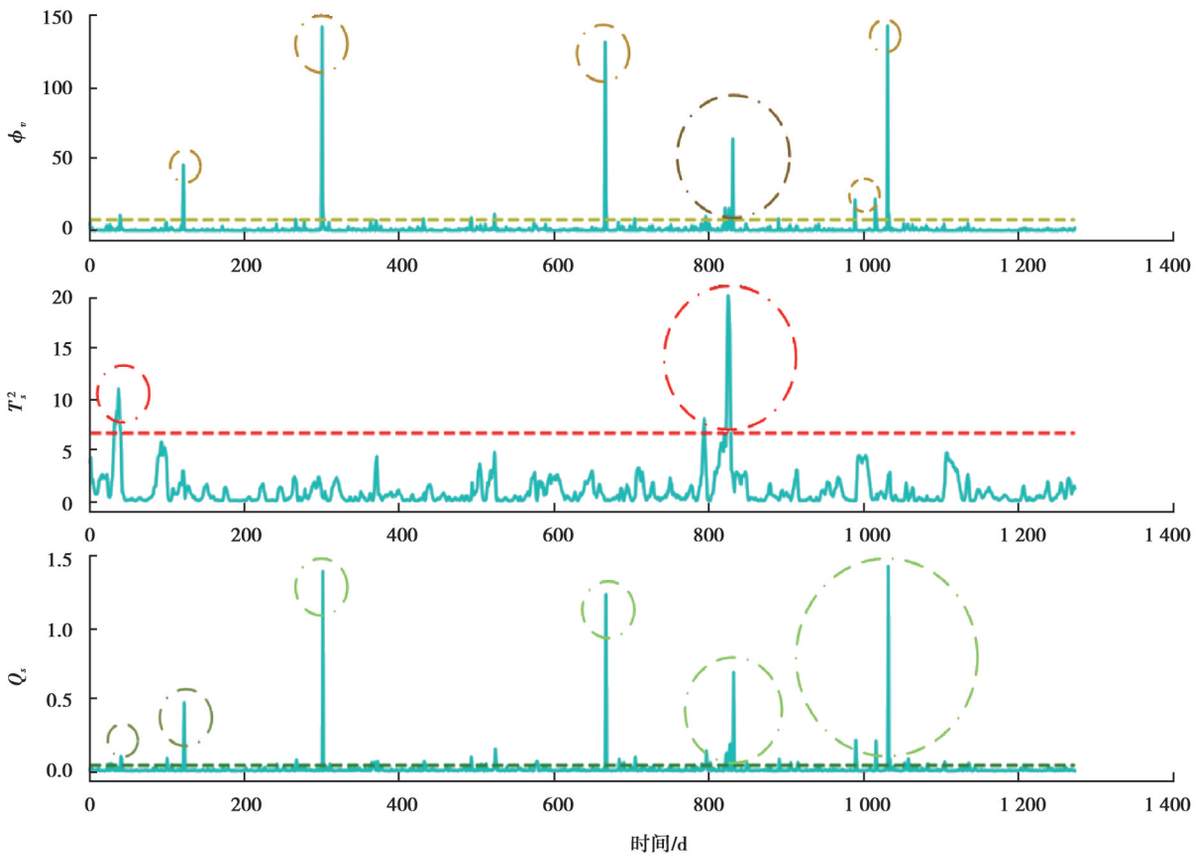

图9 DiPCA异常检测

Fig. 9 DiPCA anomaly detection

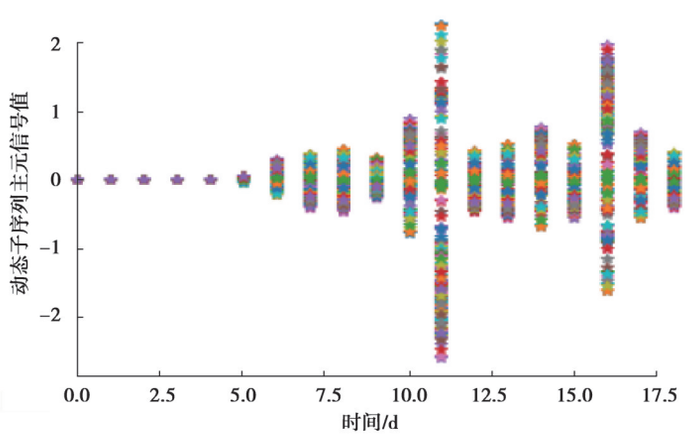

图10 动态局部子序列可视化

Fig. 10 Dynamic local subsequence visualization

图9中的指标自上向下分别为综合动态指标、霍特林统计指标、静态统计指标,三者具有相同的异常点高发范围,联系实际的空气质量状况,区分异常事件数据与错误数据。图10表示典型局部特征,其中对特征量做量纲化处理,在图中只显示了单位数值。图中以横坐标5.0为起点,数据点逐渐扩散,在此之前,同一横坐标数据集中于一点,且处于静止状态,未随时间改变。数据出现分散状态之后,均匀分布于同一时间点,不同时间点序列状态皆不相同,说明获取的特征异或性较强,动态性较高,可以有效区分序列类型。

SSA-SVM模型进行分类运算,采用惯性权重原理优化步长参数,当麻雀数量为30,精确度相比较于原算法显著提升;当麻雀数量为60,精确度已处于收敛状态并达到92%。算法召回率初期已上升至67%,后期上升趋势平稳,但最终结果未有大幅度上升,为88.6%,仍有改进前景。

根据仿真场景提供的参数,按经验缓存数据,按照任务卸载运行流程,采用SVM、朴素贝叶斯(Naive Bayesian,NB)、LSTM神经分类器、以及提出的组合算法对比不同算法的能耗、成本情况。这里的能耗是指传感设备边缘端点数据处理的能耗总和,由本地执行的CPU功率能耗以及任务卸载情况下的上传能耗与CPU空闲放电能耗共同组成;其成本费用是指简化仿真条件,只考虑任务卸载后的全体边缘设备的运营费用总和。任务数量相同的前提条件下,提出的算法通过缩小观察数据的跨度,提取实用信息,有效降低了能耗。从成本的角度看,与其他算法相比基本持平(如表4所示),可以重点考虑从优化任务卸载策略方面降低成本。

表4　不同模型的能耗、成本对比

Table 4　Comparison of energy consumption and cost of different models

算法	能耗/kW	成本/RMB
本文算法	1 647	430
SVM	1 780	480
NB	1 766	532
LSTM神经分类器	1 853	577

4　结　论

提出了一种新颖的时间序列分类组合算法,可以在压缩数据的前提下捕捉时间序列异或特征,SSA智能优化SVM算法的罚因子与核函数关键参数,充分运用动态信息,获取关键分类差异性特征,分类效果明显上升;该算法可以尝试与边缘计算框架有机结合,规避了复杂的数据重构过程,并从资源消耗的角度分析该算法实用性较高,可行性极强。但由于算法在进行数据压缩与子序列再提取的过程中,部分数据集容量较小时,可能会出现提取特征不全面的问题,这导致算法运行时间会因未满足信息增益的最低要求而延长,应思考如何调整算法的步骤使其更加适合不同特性的数据集。并在今后的研究中,重点为智能化任务卸载策略与数据分类功能结合,形成一个完整的智能框架结构。

参考文献

[1] Ku W F, Storer R H, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1995, 30(1): 179-196.

[2] 李德文, 郭胜均. 中国煤矿粉尘防治的现状及发展方向[J]. 金属矿山, 2009(S1): 747-752.
Li D W, Guo S J. Situation and development direction of dust prevention and treatment for China coal mine[J]. Metal Mine, 2009(S1): 747-752. (in Chinese)

[3] Dong Y N, Qin S J. A novel dynamic PCA algorithm for dynamic data modeling and process monitoring[J]. Journal of Process Control, 2018, 67: 1-11.

[4] 原继东, 王志海, 韩萌. 基于Shapelet剪枝和覆盖的时间序列分类算法[J]. 软件学报, 2015, 26(9): 2311-2325.
Yuan J D, Wang Z H, Han M. Shapelet pruning and shapelet coverage for time series classification[J]. Journal of Software, 2015, 26(9): 2311-2325. (in Chinese)

[5] 闫欣鸣, 孟凡荣, 闫秋艳. 基于趋势特征表示的shapelet分类方法[J]. 计算机应用, 2017, 37(8): 2343-2348, 2356.
Yan X M, Meng F R, Yan Q Y. Shapelet classification method based on trend feature representation[J]. Journal of Computer Applications, 2017, 37(8): 2343-2348, 2356. (in Chinese)

[6] Dorigo M, Maniezzo V, Colorni A. Ant system: optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 1996, 26(1): 29-41.

[7] Karaboga D, Akay B. A comparative study of artificial bee colony algorithm[J]. Applied Mathematics and Computation, 2009, 214(1): 108-132.

[8] Xue J K, Shen B. A novel swarm intelligence optimization approach: sparrow search algorithm[J]. Systems Science & Control

Engineering, 2020, 8(1): 22-34.

[9] 朱清智, 董泽, 马宁 . 基于即时学习算法的短期负荷预测方法[J]. 电力系统保护与控制, 2020, 48(7): 92-98.
Zhu Q Z, Dong Z, Ma N. Forecasting of short-term power based on just-in-time learning[J]. Power System Protection and Control, 2020, 48(7): 92-98. (in Chinese)

[10] 林昶咏, 吴桂联, 张林垚, 等 . 分布式电源接入配电系统优化规划方案[J]. 现代电力, 2019, 36(6): 82-87.
Lin C Y, Wu G L, Zhang L Y, et al. Planning scheme optimization for distributed generation accessed in distribution system[J]. Modern Electric Power, 2019, 36(6): 82-87. (in Chinese)

[11] Xiao Z W, Xu X, Zhang H X, et al. A new multi-process collaborative architecture for time series classification[J]. Knowledge-Based Systems, 2021, 220: 106934.

[12] Zhang W, Wang Z H, Yuan J D, et al. Shapelet discovery by lazy time series classification[J]. Computational Intelligence and Neuroscience, 2020, 2020: 1978310.

[13] Wang W, Hu X H, Wang M Y, et al. Implementation of parallel algorithm technology for time series data mining[J]. Journal of Physics: Conference Series, 2021,2066：012043.

[14] Arathi M. An efficient and accurate time series classification using shapelets[J]. International Journal of Information and Electronics Engineering, 2014, 4(5): 347-353.

[15] Dempster A, Petitjean F, Webb G I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels[J]. Data Mining and Knowledge Discovery, 2020, 34(5): 1454-1495.

[16] Alamir M A. A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers[J]. Applied Acoustics, 2021, 175: 107829.

[17] Zhang J T, Shen W M, Gao L, et al. Time series classification by shapelet dictionary learning with SVM-based ensemble classifier[J]. Computational Intelligence and Neuroscience, 2021, 2021: 1-13.

[18] Mahmud M A. Isolated area load forecasting using linear regression analysis: practical approach[J]. Energy and Power Engineering, 2011, 3(4): 547-550.

[19] Liang Z Y, Wang H Z. Efficient class-specific shapelets learning for interpretable time series classification[J]. Information Sciences, 2021, 570: 428-450.

[20] Zheng K D, Chen Q X, Wang Y, et al. A novel combined data-driven approach for electricity theft detection[J]. IEEE Transactions on Industrial Informatics, 2019, 15(3): 1809-1819.

[21] Chatterjee A, Siarry P. Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimization[J]. Computers & Operations Research, 2006, 33(3): 859-871.

（编辑　吕建斌）