基于 shapelets 学习的多元时间序列分类

赵慧赟 潘志松
(陆军工程大学指挥控制工程学院 南京 210007)

摘 要 多元时间序列广泛存在于日常生活中的各个领域,多元时间序列分类是从时间序列数据中获取信息的基本方法。目前,时间序列分类研究面临着相似性度量方法特殊、原始数据维度高等问题,现有的多元时间序列分类方法的分类性能仍有待提高。文中提出一种基于 shapelets 学习的多元时间序列分类方法。首先,提出了新的正则化最小二乘损失学习框架下的 shapelets 学习方法,在此基础上采用基于 shapelets 的一元时间序列分类方法对多元时间序列的每维一元数据进行分类,随后由各维上的分类结果投票决定多元时间序列的最终分类结果。实验证明,所提方法在多元时间序列分类问题中能够取得较高的分类精度。

关键词 多元时间序列,分类,shapelets,shapelets 学习
中图法分类号 TP311.11 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.05.030

Multivariate Time Series Classification Based on Shapelets Learning

ZHAO Hui-yun PAN Zhi-song
(College of Command and Control Engineering, The Army Engineering University of PLA, Nanjing 210007, China)

Abstract Multivariate time series data exist in a wide range of real-life domains, and multivariate time series classification is a basic method of obtaining information from time series data. At present, time series classification is suffered from the problem that the similarity measure of time series data is special and the dimension of the original data is high, thus the classification performance of the existing multivariate time series classification methods still need to be improved. This paper presented a multivariate time series classification method based on shapelet learning. At first, this paper established a shapelets learning method under a regularized least squares loss learning framework, and the time series classification method with one dimension based on shapelets is used to classify the vrivariate data of multivariate time series. Then the final resut of the multivariate time series is determined through plurality voting. Experimental results indicate that the proposed method achieves high classification accuracy when processing multivariate time series classification problem.

Keywords Multivariate time series, Classification, Shapelets, Shapelets learning

一个时间序列是一组序列数据,它通常是在相等间隔的时间段内,依照给定的采样率对某种潜在过程进行观测的结果[1]。时间序列数据广泛存在于现实生活中的各个领域,如股票市场、科学实验、医疗生物实验观测、传感器网络环境数据采集、移动物体位置信息服务等[2]。时间数据分为两大类:一元时间序列(Univariate Time Series, UTS)和多元时间序列(Multivariate Time Series, MTS)[3]。一个多元时间序列由多个一元时间序列组成,可理解为一次采样中能够获得不同来源的多个观测结果,多元时间序列在实际应用中更为普遍。数据分类是从实际数据中获取信息的基本手段,因此近年来多元时间序列分类(Multivariate Time Series Classification, MTSC)问题得到研究者的广泛关注。

目前,时间序列分类研究面临着许多困难。一方面,分类问题的本质是数据相似性度量,例如,在最近邻算法中,通过计算待分类样本与已有类别标记样本之间的距离来判定待分类样本的类别,然而由于时间序列数据本身具有时间先后顺序,且不同时间序列之间可能存在相位差,一般的基于欧氏距离的相似性度量方法不能很好地描述时间序列数据之间的相似性。另一方面,时间序列数据的维度普遍较高,直接对原始多元时间序列数据进行分类通常需要消耗大量的计算资源[4]。

基于 shapelets 的时间序列分类方法是有效解决上述问题的方法之一。shapelets 是指一个时间序列中最具辨识性的子序列(见图 1),这一概念由 Keogh 等于 2009 年首次提出[5]。最早的基于 shapelets 的时间序列分类方法以 Information Gain(IG)等评价指标为依据选定最优 shapelets,而后以最优 shapelets 为节点构建决策树来对时间序列进行分类[5]。Hills 等于 2014 年提出 shapelets transformation 概念,首先在训练数据中查找或学习 K 个 shapelets,然后计算原始

到稿日期:2017-11-08 返修日期:2018-02-21 本文受国家自然科学基金(61473149)资助。
赵慧赟(1990—),女,博士生,主要研究方向为机器学习及优化方法在时间序列分析中的应用,E-mail:zhaohuiyun1819@126.com;**潘志松**(1973—),男,教授,博士生导师,主要研究方向为模式识别、机器学习、网络安全,E-mail:hotpzs@hotmail.com(通信作者)。

时间序列到 K 个 shapelets 之间的距离，并将这些距离作为当前时间序列的特征，以此将原始时间序列映射到新特征空间，新特征空间中的数据可以用 SVM 等一般分类器进行分类[6]。一般分类器能够对通过 shapelets transformation 转换后的时间序列数据进行分类，大大提高了时间序列的分类效率，因此后续基于 shapelets 的时间序列分类研究大都以此为基础展开。基于 shapelets 和 shapelets transformation 的时间序列分类方法利用 shapelets 来度量各个时间序列之间的相似性。查找 shapelets 的过程也是一种模式识别过程。同时，由于最优 shapelets 的个数少于原始时间序列的维数，因此新特征空间中的样本维数也相应降低。目前，研究者们对基于 shapelets 的一元时间序列分类问题已经进行了比较深入的研究[7-10]，但仍面临着查找 shapelets 非常耗时的问题。在现有研究的基础上，一个简单的想法是：将基于 shapelets 的一元时间序列分类方法应用于多元时间序列分类。

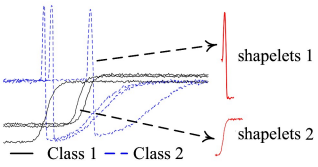


图 1　shapelets 示意图

Fig. 1　Schematic diagram of shapelets

因此，针对多元时间序列分类问题，本文提出了一种基于 shapelets 学习的多元时间序列分类方法（Multivariate Time Series Classification Method Based on Shapelets Learning，SL_{MTSC}）。首先建立了新的一元 shapelets 学习框架，针对多元时间序列中的每一维时间序列，建立正则化最小二乘损失目标函数，通过一种加速 Nesterov 优化方法 EFLA[11] 来求解目标函数以获取 shapelets；然后对一维数据进行特征空间转换并采用 SVM 进行分类。该学习框架大大提高了基于 shapelets 的一元时间序列的分类效率。在获取多个一元时间序列分类结果的基础上，通过投票决定一个多元时间序列的最终类别。实验证明，所提方法对于多元时间序列具有较高的分类精度。

1　相关研究

近年来，针对多元时间序列分类的问题已取得许多研究成果，现有研究主要以时间序列特征选择为入手点，通过设计适用于多元时间序列的特征选择方法来选择具有代表性的特征，所选特征构成新数据，可采用现有针对一般数据的分类器对新数据进行分类。Yoon 等[12] 提出了一种基于主成分分析（PCA）的多元时间序列特征子集选择方法。Li 等[13-14] 在 2006 年和 2007 年相继提出了两种基于奇异值分解（SVD）的特征选择方法以用于多元时间序列分类。He 等[15] 于 2015 年提出了一种 MCFEC 方法，从多元时间序列样本中获取核心特征，并在此基础上提出 MCFEX-rule 和 MCFEC-QBC 两种方法，以用于多元时间序列分类。Li 等[16] 于 2016 年提出了一种基于公共主成分分析的方法，用于降低多元时间序列的数据复杂度，并在此基础上完成分类。上述方法的主要问题在于在特征选择的过程中或多或少地损失了时间序列数据本身的时序性特征。为了保持时间序列的固有属性，Górecki[17] 采用传统的动态时间规整（DTW）方法、微分动态时间规整（DDTW）方法及扩展微分动态时间规整（DD_{DTW}）方法来度量多元时间序列的相似性，并采用最近邻法进行分类。Wang 等[3] 直接利用递归神经网络（RNN）方法来进行多元时间序列分类，并采用自适应差分进化算法（ADE）进行参数调整。实验结果证明，上述两种方法具有较高的分类精度，但 DTW 计算过程及神经网络的训练过程比较耗时。

查找 shapelets 的过程在某种程度上也是一种特征选择过程，但 shapelets 能够较好地保存时间序列原有的时间先后顺序。因此，已有许多研究者将基于 shapelets 的一元时间序列分类方法扩展至多元时间序列分类中。Mueen 等[18] 提出一个简单的想法，将多元时间序列按序拼接为一个一元时间序列，然后采用所提出的 fast shapelets 方法进行分类。2012 年，Ghalwash 等[19] 设计了多元 shapelets 检测法来查找多元时间序列中的 shapelets 并完成分类。基于 shapelets 的时间序列分类方法面临的最大问题是查找时间序列非常耗时，因此 Rakthanmanon 提出在多元时间序列的各个维度上随机地选取 shapelets 用以完成分类[20]，该方法有效提高了多元时间序列分类的效率，但分类精度相比其他方法略低。

综上所述，现有多元时间序列分类方法在分类效率和精度上仍有待提高，因此本文提出一种新的基于 shapelets 学习的多元时间序列分类方法以解决上述问题。

2　基于 shapelets 学习的多元时间序列分类方法

Dietterich 等证明，对一组决策树利用投票方式进行组合能够有效减小分类器偏置[21]。同理，我们假设组成多元时间序列的每一维相互独立，那么通过首先对每一维上的一元时间序列进行分类，再用投票方式决定多元时间序列类别的方法，能够有效解决多元时间序列分类的问题。本文提出的基于 shapelets 学习的多元时间序列分类方法的整体架构如图 2 所示。训练过程中，单独训练多元时间序列的每一维，通过基于 shapelets 的一元时间序列分类方法训练分类器。对于测试数据，首先采用训练好的分类器对每一维时间序列进行分类，随后各个维的分类器投票决定一个多元时间序列的最终类别。

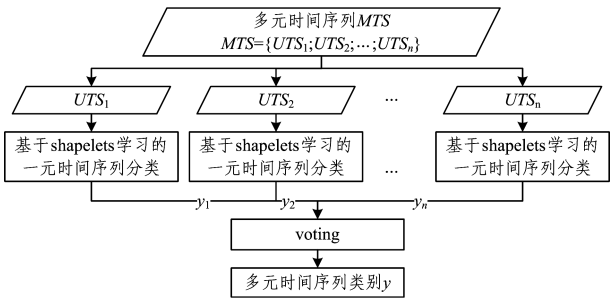


图 2　基于 shapelets 学习的多元时间序列分类方法的整体架构

Fig. 2　Framework of the proposed shapelets-based MTSC method

2.1　基本定义

定义 1（时间序列）　假设多元时间序列数据集 MTS 中包含 N 个样本，每个样本为 $MTS_i \in \mathbb{R}^{d\times Q}$，$i \in \{1,2,\cdots,N\}$，$d$ 为每个样本包含的一元时间序列的个数，Q 为单个样本一

维的长度。将数据集中每个样本的第 j 维($j\in\{1,2,\cdots,d\}$)数据组成一个一元时间序列数据集,记为 $\boldsymbol{X}^j$。为方便讨论,2.2 节直接标记一个一元时间序列数据集为 $\boldsymbol{X}$,且 $\boldsymbol{X}=\{(\boldsymbol{X}_1,y_1),(\boldsymbol{X}_2,y_2),\cdots,(\boldsymbol{X}_N,y_N)\}$。其中,单个样本 $\boldsymbol{X}_i\in\mathbb{R}^Q$,样本类标 $y_i\in\{1,2,\cdots,C\}$,$i\in\{1,\cdots,N\}$。

定义 2(shapelets) shapelets 是时间序列中能够最大程度区分不同类别时间序列的子序列。在一元时间序列分类问题中选出 K 个最优 shapelets,并记为 $\boldsymbol{S}\in\mathbb{R}^{K\times M}$(其中 M 为 shapelets 的长度,实际学习到的 shapelets 长度均不同,为了便于说明,统一记为 M)。一个长度为 M 的 shapelets $\boldsymbol{S}_k=\{s_{k1},s_{k2},\cdots,s_{kM}\}$($k\in\{1,\cdots,K\}$)是时间序列 $\boldsymbol{X}_i=\{x_{i1},x_{i2},\cdots,s_{k1},s_{k2},\cdots,s_{kM},\cdots,x_{iQ}\}$($i\in\{1,\cdots,N\}$,$k\in\{1,\cdots,K\}$,$M<Q$)的子序列。

定义 3(时间序列与 shapelets 之间的距离) 第 i 个时间序列 $\boldsymbol{X}_i$ 与第 k 个 shapelets $\boldsymbol{S}_k$ 之间的距离如式(1)所示。其表示依次计算 $\boldsymbol{S}_k$ 与 $\boldsymbol{X}_i$ 的所有 $J(J:=Q-M+1)$个长度为 M 的子序列间的欧氏距离,选择其中最小的距离作为 $\boldsymbol{X}_i$ 与 $\boldsymbol{S}_k$ 之间的距离。

$$D_{i,k}=\min_{j=1,\cdots,J}\frac{1}{M}\sum_{m=1}^{M}(X_{i,j+m-1}-S_{k,m})^2 \tag{1}$$

定义 4(shapelets transformation[6]) 在找到 K 个最优 shapelets 后,根据定义 3,计算 K 个最优 shapelets 与一个时间序列间的距离作为该时间序列的新特征,时间序列样本集被映射至新特征空间,这一过程被称为 shapelets transformation。数据集 $\boldsymbol{X}$ 转换为 $\boldsymbol{D}=\{(\boldsymbol{D}_1,y_1),(\boldsymbol{D}_2,y_2),\cdots,(\boldsymbol{D}_N,y_N)\}$,其中 $\boldsymbol{D}_i=\{D_{i,1},D_{i,2},\cdots,D_{i,k},\cdots,D_{i,K}\}$。由于 $K<Q$,shapelets transformation 的过程降低了时间序列数据的维度,同时新样本集可以用一般分类器进行分类。

2.2 基于 shapelets 学习的一元时间序列分类

由图 2 可知,基于 shapelets 学习的一元时间序列分类是本文方法的关键步骤。自 Hills 等[6]提出 shapelets transformation 的概念后,一般的基于 shapelets 的分类方法的流程如图 3 所示。

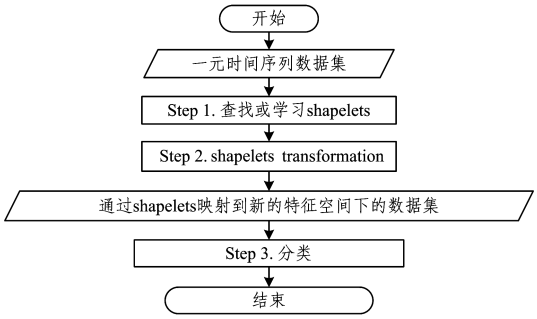


图 3 基于 shapelets 的时间序列分类的一般流程

Fig. 3 General flow chart of shapelets-based TSC method

首先通过遍历或学习数据集找到最优的 shapelets,然后计算原始时间序列与各个 shapelets 之间的距离,以这些距离作为时间序列的新特征,并将时间序列映射至新特征空间,最后用一般分类器完成分类。本节针对查找 shapelets 这一关键步骤,提出了正则化最小二乘目标函数,然后通过 EFLA 求解优化目标从而获得 shapelets,避免了遍历数据集,极大地提高了 shapelets 查找的效率。对于新特征空间下的数据,采用支持向量机(SVM)进行分类。

2.2.1 目标函数

Fused Lasso 是近年来常用的稀疏学习方法,Fused Lasso的正则化项表达式为:

$$\lambda_1\|\boldsymbol{w}\|_1+\lambda_2\|R\boldsymbol{w}\|_1,\boldsymbol{w}\in\boldsymbol{R}^{n\times 1} \tag{2}$$

其中,$R\in\mathbb{R}^{(n-1)\times n}$是一个稀疏矩阵,定义为:

$$R_{ij}=\begin{cases}-1, & j=i,i=1,2,\cdots,n-1\\ 1, & j=i+1,i=1,2,\cdots,n-1\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

Fused Lasso 正则化项保证了解向量在相邻位置的值尽可能相近,从而保留了问题本身所包含的前后顺序信息,同时能够确保解向量的稀疏性。因此,包含 Fused Lasso 正则化项的优化问题,其解向量表现出成块稀疏特性。

在分类和回归问题中,最小二乘损失是常见的损失函数之一,如式(4)所示。

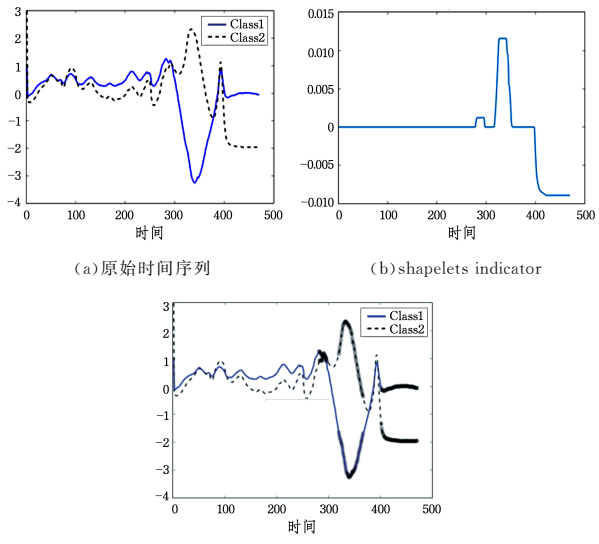
$$\boldsymbol{w}=\arg\min_{\boldsymbol{w}}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 \tag{4}$$

满足最小二乘损失约束的解向量能够最大化样本类间的间隔。同时,最小二乘损失具有贝叶斯一致性,当样本量趋于无穷时,包含最小二乘损失的分类器的分类准确率无限接近于贝叶斯分类器。换言之,采用最小二乘损失能够从理论上保证分类面最优。

通过结合 Fused Lasso 正则化项和最小二乘损失函数,本文构建了新的 shapelets 学习优化目标,如式(5)所示。

$$\min_{\boldsymbol{w}}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2+\lambda_1\|\boldsymbol{w}\|_1+\lambda_2\|R\boldsymbol{w}\|_1 \tag{5}$$

其中,$\boldsymbol{w}$ 为解向量。在 Fused Lasso 正则化项和最小二乘损失约束下,解向量是一个能够最大程度区分不同类别样本的成块稀疏向量,即 $\boldsymbol{w}$ 是由有限 0 块和非 0 块组成的向量。根据 shapelets 的定义,$\boldsymbol{w}$ 中非 0 块所在位置对应的子序列就是待查找的 shapelets,因此将该向量定义为 shapelets indicator。原始时间序列、shapelets indicator 和 shapelets 之间的关系如图 4 所示,图 4(c)中的加粗部分即为所找到的 shapelets。



(a)原始时间序列 (b)shapelets indicator

(c)shapelets

图 4 原始时间序列、shapelets indicator 与 shapelets 之间的关系示意图

Fig. 4 Relationship among time series, shapelets indicator and shapelets

2.2.2　EFLA

本文采用 EFLA 方法[11]来求解上述 shapelets 学习优化目标。首先，优化目标式(5)可以改写为：

$$\min_{\boldsymbol{w}} h(\boldsymbol{w})=loss(\boldsymbol{w})+fl(\boldsymbol{w}) \tag{6}$$

其中，$loss(\boldsymbol{w})=\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2$，$fl(\boldsymbol{w})=\lambda_1\|\boldsymbol{w}\|_1+\lambda_2\|R\boldsymbol{w}\|_1$。对 $loss(\cdot)$在 $\boldsymbol{w}$ 进行一阶泰勒展开后得：

$$h_{L,w}(\boldsymbol{v})=[loss(\boldsymbol{w})+\langle loss'(\boldsymbol{w}),\boldsymbol{v}-\boldsymbol{w}\rangle]+fl(\boldsymbol{v})+\frac{L}{2}\|\boldsymbol{v}-\boldsymbol{w}\|^2 \tag{7}$$

正则化项$(L/2)\|\boldsymbol{v}-\boldsymbol{w}\|^2(L>0)$确保 $\boldsymbol{v}$ 不会远离 $\boldsymbol{w}$，因此式(7)近似于 $h(\cdot)$。然后用 EFLA 算法求解式(7)。

EFLA 算法如算法 1 所示，其中步骤 4 是关键步骤，式(8)被用于迭代求解式(7)：

$$\boldsymbol{w}_{i+1}=\arg\min_{\boldsymbol{v}} h_{L_i,\boldsymbol{w}_i}(\boldsymbol{v}) \tag{8}$$

其中，$\{L_i\}$是一系列根据线搜索得到的 Lipschitz 常数。

算法 1　EFLA[11]

Input：$\lambda_1\geqslant 0,\lambda_2\geqslant 0,L_0>0,\mathbf{w}_0,T$

Output：$\mathbf{w}_{T+1}$

1. Initialize $\mathbf{w}_1=\mathbf{w}_0,\alpha_{-1}=0,\alpha_0=1,L=L_0$
2. for i=1 to T do
3. set $\beta=(\alpha_{i-2}-1)/\alpha_{i-1},\mathbf{p}_i=\mathbf{w}_i+\beta_i(\mathbf{w}_i-\mathbf{w}_{i-1})$
4. find $\mathbf{w}_{i+1}=\arg\min_{\mathbf{v}} h_{L_i,\mathbf{w}_i}(\mathbf{v})$ with Fused Lasso Signal Approximator (FLSA)
5. find the smallest $L=L_{i-1},2L_{i-1},\cdots$ such that $h(\mathbf{w}_{i+1})\leqslant h_{L,\mathbf{p}_i}(\mathbf{w}_{i+1})$
6. set $L_i=L$ and $\alpha_i=\alpha_{i-1},\alpha_{i+1}=(1+\sqrt{1+4\alpha_i^2})/2$
7. end for

EFLA 算法第 4 步中的 FLSA 方法通常用于求解以下问题：

$$\min_{\boldsymbol{w}\in R^n} f_{\lambda_2}^{\lambda_1}(\boldsymbol{w})\equiv\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{q}\|^2+\lambda_1\|\boldsymbol{w}\|_1+\lambda_2\|R\boldsymbol{w}\|_1 \tag{9}$$

令 $\pi_{\lambda_2}^{\lambda_1}(q)\equiv\arg\min_{\boldsymbol{w}} f_{\lambda_2}^{\lambda_1}(\boldsymbol{w})$，当 $\boldsymbol{q}=\boldsymbol{p}_i-loss'(\boldsymbol{p}_i)/L_i$ 时，可以证明：

$$\pi_{\lambda_2}^{\lambda_1}(\boldsymbol{q})=\arg\min_{\boldsymbol{v}} h_{L_i,\boldsymbol{p}_i}(\boldsymbol{v}) \tag{10}$$

因此，FLSA 方法求出的式(10)的解即为 EFLA 算法第 4 步的解 $\boldsymbol{w}_{i+1}$。文献[22]证明，对于任意(λ_1,λ_2)，式(9)的解可以通过参数为$(0,\lambda_2)$时的求解结果计算获得，如定理 1 所示。

定理 1[20]　对于任意 $\lambda_1,\lambda_2\geqslant 0$，有：

$$\pi_{\lambda_2}^{\lambda_1}(\boldsymbol{q})=\mathrm{sgn}(\pi_{\lambda_2}^0(\boldsymbol{q}))\odot\max(|\pi_{\lambda_2}^0(\boldsymbol{q})|-\lambda_1,0) \tag{11}$$

因此，只需求出当 $\lambda_1=0$ 时式(9)的解即可，即：

$$\min_{\boldsymbol{w}\in R^n} f_{\lambda_2}^{\lambda_1}(\boldsymbol{w})\equiv\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{q}\|^2+\lambda_2\|R\boldsymbol{w}\|_1 \tag{12}$$

对于式(12)，引入对偶变量 $\boldsymbol{z}\in\mathbb{R}^{n-1}$，得到问题(12)的对偶问题：

$$\min_{\|\boldsymbol{z}\|_\infty\leqslant\lambda_2}\varphi(\boldsymbol{z})\equiv\frac{1}{2}\|R^{\mathrm{T}}\boldsymbol{z}\|^2-\langle R^{\mathrm{T}}\boldsymbol{z},\boldsymbol{q}\rangle \tag{13}$$

为求解式(13)，首先通过 Rose 算法(见算法 2)计算线性系统 $RR^{\mathrm{T}}\boldsymbol{z}=R\boldsymbol{q}$ 的解 $\hat{\boldsymbol{z}}$，并获取 $\lambda_2^{\max}=\|\hat{\boldsymbol{z}}\|_\infty$。当 $\lambda_2\geqslant\lambda_2^{\max}$ 时，$\hat{\boldsymbol{z}}$即为式(13)的解；当 $0<\lambda_2<\lambda_2^{\max}$ 时，采用 SFA 算法(见算法 3)求解式(13)。求得式(13)的最优解 $\boldsymbol{z}^*$ 后，计算$\pi_{\lambda_2}^0(\boldsymbol{q})=\boldsymbol{q}-R^{\mathrm{T}}\boldsymbol{z}^*$，再由定理 1 求得 $\pi_{\lambda_2}^{\lambda_1}(\boldsymbol{q})$。至此，EFLA 算法的关键步骤执行完毕。

算法 2　Rose[23]

Input：$\mathbf{u}\in\mathbb{R}^{(Q-1)\times 1}$

Output：$\hat{\mathbf{z}}\in\mathbb{R}^{(Q-1)\times 1}$ satisfying $RR^{\mathrm{T}}\mathbf{z}=\mathbf{u}$

1. Compute the scalar $\mathbf{s}=-Q^{-1}\sum_{j=1}^{Q-1} j\times\mathbf{u}_j$
2. Compute $\hat{\mathbf{z}}_j$ sequentially using $\hat{\mathbf{z}}_{Q-1}=\mathbf{u}_{Q-1}+\mathbf{s}$ and $\hat{\mathbf{z}}_j=\hat{\mathbf{z}}_{j+1}+\mathbf{u}_j,j=Q-2,\cdots,1$
3. Obtain $\hat{\mathbf{z}}_j$ sequentially using $\hat{\mathbf{z}}_j=\hat{\mathbf{z}}_{j+1}+\mathbf{u}_j,j=2,\cdots,Q-1$

算法 3　SFA[11]

Input：$\mathbf{p}\in\mathbb{R}^{(Q-1)\times 1},0<\lambda_2<\lambda_2^{\max},T_3$

Output：$\mathbf{z}_k\in\mathbb{R}^{(Q-1)\times 1}$ satisfying $RR^{\mathrm{T}}\mathbf{z}=\mathbf{u}$

1. Compute $\mathbf{z}_0\in\mathbb{R}^{(Q-1)\times 1}$ through the restart technique proposed in reference [11];
2. Set $L=2-2\cos(\pi(Q-1)/Q)$
3. for i=1 to T_3 do
4. Compute $\mathbf{g}_i=\varphi'(\mathbf{z}_i)=RR^{\mathrm{T}}\mathbf{z}_i-R\mathbf{q}$
5. Set $z_{i+1}=P_{\lambda_2}(z_i-g_i/L)$
6. end for

2.2.3　一元时间序列分类

采用 EFLA 算法求解出 shapelets indicator $\boldsymbol{w}$ 后，通过 $\boldsymbol{w}$ 在原始时间序列中定位 shapelets。假设 $\boldsymbol{w}=[0,\cdots,0,w_{t_{is}},\cdots,w_{t_{ie}},0,\cdots,0,w_{t_{N_{is}}},\cdots,w_{t_{N_{ie}}},0,\cdots,0]$中共有 N_B 个非零块，其中第 i 个非零块始于位置 t_{is} 并在 t_{ie} 处结束。原始时间序列 $\boldsymbol{X}=\{(\boldsymbol{X}_1,y_1),(\boldsymbol{X}_2,y_2),\cdots,(\boldsymbol{X}_N,y_N)\}$，那么通过 $\boldsymbol{w}$ 定位到的所有 shapelets 为：

$$\boldsymbol{S}=\{[\boldsymbol{X}_k]_{t_{is}:t_{ie}}:=k=1,2,\cdots,N,i=1,2,\cdots,N_B\} \tag{14}$$

其中，$[\boldsymbol{X}_k]_{t_{is}:t_{ie}}$是样本 $\boldsymbol{X}_k$ 从位置 t_{is} 到 t_{ie} 的子序列。shapelets 的数目$|\boldsymbol{S}|=N_B\times N$。

求得 shapelets 后，通过 2.1 节中的式(1)计算所有时间序列到各个 shapelets 之间的距离，将$|\boldsymbol{S}|$个距离作为原始时间序列的新特征。利用新特征空间下的数据训练 SVM 分类器。

2.3　多元时间序列分类

对于多元时间序列数据集 $MTS\in\mathbb{R}^{d\times N}$ 中的每个一元时间序列训练数据子集 $X^i(i\in\{1,2,\cdots,d\})$，采用 2.2 节的一元时间序列分类方法学习 shapelets 并构建 SVM 分类器。假设一个测试样本 $MTS_i(i\in\{1,2,\cdots,N\})$各维上的一元时间序列的分类结果为$\{y^1,y^2,\cdots,y^d\}$，采用相对多数投票法(Plurality Voting)投票决定 MTS_i 的最终分类结果 y，即统计$\{y^1,y^2,\cdots,y^d\}$的众数作为 MTS_i 的分类结果，当众数为 0 或者多个时，从中随机选取一个值作为 MTS_i 的分类结果。

3　多元时间序列分类实验

为了分析本文提出的基于 shapelets 学习的多元时间序列分类方法的有效性，设计了多元时间序列分类实验，将本文方法 SL_{MTSC} 与 DTW、DDTW、文献[16]中的 DD_{DTW} 以及文献[3]中的 C_{ADE} 等分类方法进行比较。

实验使用 MATLAB 2015a 进行数值计算，SVM 的实现采用了 Libsvm 软件包。实验环境为普通台式电脑，操作系统为 Windows 7 旗舰版，其中 CPU 为 Intel i5 处理器，内存为 16 GB。

3.1 数据集描述

实验在 10 个数据集上展开，为便于比较，选取文献[3]所使用的数据集。数据集的类别、维数、样本个数及数据来源等如表 1 所列。

表 1 数据集的详细信息

Table 1 Details of datasets

数据集名称	类别数	维数	样本长度	样本个数	数据来源
uWaveGestureLibrary	8	3	315	4478	UCR
ECG	2	2	152	200	Olszewski
Wafer	2	6	198	1194	Olszewski
Libras	15	2	45	360	UCI
Pendigits	10	2	8	10992	UCI
Robot failure LP1	4	6	15	88	UCI
Robot failure LP2	5	6	15	47	UCI
Robot failure LP3	4	6	15	47	UCI
Robot failure LP4	3	6	15	117	UCI
Robot failure LP5	5	6	15	164	UCI

3.2 实验结果及分析

实验统计了 5 种方法在 10 个多元时间序列数据集上的分类错误率，结果如表 2 所列(其中粗体表示每个数据集上分类错误率最低的实验结果；括号中的数据表示 5 种方法的分类错误率由低到高的排序，错误率最低的排序为 1)。从表 2 可以看出，本文提出的 SL_{MTSC} 方法在其中 5 个数据集上的分类结果优于其他方法，证明了所提方法的有效性。

表 2 多元时间序列分类的错误率

Table 2 Error rate of different MTSC methods

(单位：%)

数据集	C_{ADE}	DTW	DDTW	DD_{DTW}	SL_{MTSC}
uWaveGestureLibrary	1.92(3)	1.90(2)	3.55(4)	**1.50(1)**	4.22(5)
ECG	10.00(2)	18.50(5)	14.00(3)	14.50(4)	**7.89(1)**
Wafer	1.73(2)	2.01(4)	9.21(5)	1.92(3)	**0.52(1)**
Libras	**4.16(1)**	8.61(5)	4.17(2)	5.00(4)	4.7(3)
Pendigits	1.56(4)	0.65(3)	0.61(2)	**0.50(1)**	1.57(5)
Robot failure LP1	**0.75(1)**	12.64(3)	22.50(5)	14.86(4)	6.06(2)
Robot failure LP2	**23(1)**	32(2)	38(4)	32(2)	33.33(3)
Robot failure LP3	22.86(2)	29(4)	29(4)	25(3)	**19.23(1)**
Robot failure LP4	2.85(2)	10.08(3)	20.45(4)	10.08(3)	**2.44(1)**
Robot failure LP5	27.35(2)	29.30(4)	37.32(5)	28.75(3)	**25.26(1)**
Total win	3	0	0	2	**5**

此外，本文采用相对错误率来评价所提方法的有效性。相对错误率是一种衡量各种方法相对性能的指标[24]。对于两种方法 A 和 B(其中 A 是基准方法，B 是待比较方法)，其分类错误率分别记为 ε_A 和 ε_B，则方法 B 相对方法 A 的相对错误率为：

$$\frac{\varepsilon_B-\varepsilon_A}{\varepsilon_A} \tag{15}$$

根据相对错误率的定义可知，当相对错误率为负时，待比较方法有效地提高了对问题的分类精度。令本文提出的 SL_{MTSC} 方法为待比较方法，其余 4 种方法为基准方法，根据表 2 及式(15)计算提出的 SL_{MTSC} 方法与其他 4 种方法的相对错误率，结果如图 5 所示。从图中可以看出，与其他 4 种方法相比，本文提出的方法在大部分数据集上的相对错误率为负数。在个别数据集如 LP1 中，SL_{MTSC} 方法的分类性能差于 C_{ADE} 方法，这一现象说明很少有普适的分类器对所有数据都能获得极高的分类精度。总体而言，SL_{MTSC} 方法具有较高的分类精度。

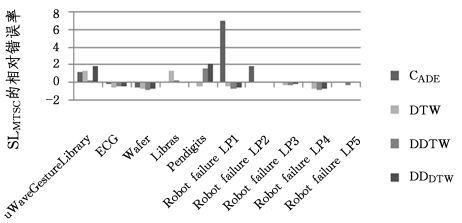


图 5 SL_{MTSC} 方法与其他 4 种方法的相对错误率的比较

Fig. 5 Relative error rate of SL_{MTSC} compared with four methods

在耗时方面，DTW、DDTW 和 DD_{DTW} 方法均采用了动态时间规整来度量时间序列之间的相似性，并用 1-NN 方法进行了分类，当训练集有 N 个长度为 Q 的 d 元时间序列时，用 DTW 等方法完成分类的时间复杂度为 $O(dNQ^2)$。相比 C_{ADE} 和本文提出的 SL_{MTSC} 方法，其他 3 种方法的时间复杂度较低，但其分类精度低于 C_{ADE} 和 SL_{MTSC}。C_{ADE} 方法利用 RNN 实现分类，通常 RNN 方法需要较长的训练时间来提高分类精度。而本文提出的 SL_{MTSC} 方法能够在秒级时间内完成多元时间序列数据集上的训练和分类。总体而言，本文提出的 SL_{MTSC} 方法针对多元时间序列具有较好的分类性能。

结束语 本文提出一种基于 shapelets 学习的多元时间序列分类方法。首先针对一元时间序列分类设计新的 shapelets 学习方法，通过求解正则化最小二乘损失目标函数来获得 shapelets，计算 shapelets 与原时间序列间的距离，并将原时间序列映射到新特征空间，采用 SVM 分类器对新特征空间的数据进行分类。在获得一个多元时间序列各个维度上的一元时间序列的分类结果后，通过相对多数投票方式来决定多元时间序列最终的分类结果。实验结果表明，本文方法在大部分多元时间序列数据集上具有较好的分类精度。

本文通过一种简单的 ensemble 方法，在求解一元时间序列分类问题的基础上解决多元时间序列分类问题，并取得了良好的分类效果，但该方法忽略了多元时间序列每维一元序列之间的相关性。如何在考虑多元时间序列内部相关性的条件下完成多元时间序列分类将是下一步研究的重点。

参 考 文 献

[1] YUAN J D, WANG Z H. Review of Time Series Representation and Classification Techniques[J]. Computer Science, 2015, 42(3): 1-7. (in Chinese)
原继东，王志海. 时间序列的表示与分类算法综述[J]. 计算机科学，2015，42(3)：1-7.

[2] LI Z X, ZHANG F M, ZHANG X F, et al. Survey of similarity search for multivariate time series[J]. Control and Decision, 2017, 32(4): 577-583. (in Chinese)
李正欣，张凤鸣，张晓丰，等. 多元时间序列相似性搜索研究综述[J]. 控制与决策，2017，32(4)：577-583.

[3] WANG L, WANG Z, LIU S. An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm[J]. Expert Systems with Applications, 2016, 43(C): 237-249.

[4] DING H, TRAJCEVSKI G, SCHEUERMANN P, et al. Querying and mining of time series data: experimental comparison of representations and distance measures[J]. Proceedings of the VLDB Endowment, 2008, 1(2): 1542-1552.

(下转第 219 页)