控制与决策
Control and Decision

基于关键形态特征的多元时间序列降维方法

李海林， 梁叶

引用本文:
李海林, 梁叶. 基于关键形态特征的多元时间序列降维方法[J]. 控制与决策, 2020, 35(3): 629−636.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.0750

文章编号: 1001-0920(2020)03-0629-08 DOI: 10.13195/j.kzyjc.2018.0750

基于关键形态特征的多元时间序列降维方法

李海林†, 梁 叶

(1. 华侨大学 工商管理学院，福建 泉州 362021；2. 华侨大学 应用统计与大数据研究中心，福建 厦门 361021)

摘 要: 针对传统主成分分析及相关方法对多元时间序列特征表示的局限性,以及降维效果对数据相似性度量质量的影响,从数据形态特征的角度出发,提出一种关键形态特征的多元时间序列降维方法. 利用动态时间弯曲方法找出训练集每个类别的中心多元时间序列,根据形态特征找出每个中心多元时间序列的关键特征变量分量的重要度,使用重要度提取若干个关键特征变量分量,达到数据降维的目的. 实验结果表明,与传统方法相比,所提方法能够有效地根据形态特征对多元时间序列进行降维,并且能够取得更好的分类效果.

关键词: 数据降维；多元时间序列；动态时间弯曲；形态特征；分类

中图分类号: TP273 **文献标志码:** A

Dimension reduction for multivariate time series based on crucial shape features

LI Hai-lin†, LIANG Ye

(1. School of Business Management，Huaqiao University，Quanzhou 362021，China；2. Research Center for Applied Statistics and Big Data，Huaqiao University，Xiamen 361021，China)

Abstract: Principal component analysis and relevant techniques are often used to represent feature for multivariate time series. However, they have some limitations for feature representation, and the dimension reduction results impact on the similarity measure accuracy. Therefore, from the perspective of shape features, a crucial shape feature dimension reduction technique for multivariate time series is proposed. In train datasets, central for multi variate time series of each category is obtained through dynamic time warping, and the importance degree of crucial feature properties components is found according to shape feature. In this way, the dimensionality of multi-time series can be reduced and the original data can be represented by the crucial shape feature. The experiments results show that the proposed method is superior because it can obtain a better classification effect and effective dimension reduction.

Keywords: data dimension reduction；multivariate time series；dynamic time warping；shape feature；classification

0 引 言

多元时间序列数据广泛存在于现实生活中,如金融股票每日的开收盘、成交量等数据,工业精密仪器设备运行状态数据,生物医学实验中各种测量数据如脑电波数据,以及气象信息数据等. 随着多元时间序列获取技术的不断发展,多元时间序列数据挖掘得到越来越多研究者的关注. 由于多元时间序列数据的高维性影响着其数据挖掘的过程和质量,多元时间序列成为数据挖掘领域的重要挖掘对象之一.

时间序列数据挖掘的任务主要有分类、聚类、相似性度量、模式识别、关联规则、异常检测等,这些挖掘任务的效率及效果通常受到时间序列数据的特征及其复杂程度的影响. 为了提高多元时间序列数据挖掘技术的性能,通常先利用数据降维和特征表示方法来降低多元时间序列数据的复杂性,进而在低维空间展开数据挖掘工作. 目前,已经有不少多元时间序列数据降维及特征表示的相关方法,如主成分分析[1-3]、奇异值分解[4-5]、独立成分分析[6-7]等方法. 其中,主成分分析方法是多元时间序列数据挖掘技术中重要的降维方法,通过数据坐标变换将原始数据以反映数据特征分布的若干个成分进行表示,进而达到将数据从高维空间映射到低维空间的目的.

时间序列形态特征可以较为客观地反映其变化的趋势,利用形态特征提取时间序列数据信息可

收稿日期: 2018-06-01；**修回日期:** 2018-10-08.

基金项目: 国家自然科学基金项目(71771094, 61300139)；福建省自然科学基金项目(2019J01067)；福建省高等学校新世纪优秀人才支持计划项目(Z1625112).

责任编委: 阳春华.

†通讯作者. E-mail: hailin@mail.dlut.edu.cn.

以为后期数据挖掘工作提供可靠的保障[8]. 传统主成分分析方法通过计算两条时间序列的协方差来衡量两条时间序列的相关性,并利用方差贡献率的大小进行选择成分,并不能很好地反映时间序列各个分量形态特征上的差异性. 同时,计算两条多元时间序列的欧氏距离只能从时间对应的顺序上来体现两者的整体关系,未能很好地反映时间序列本身的局部形态变化[9]. 降维后的数据仍然存在长度不同的问题,需要能够度量不等长度时间序列的度量方法来解决问题,如动态时间弯曲(dynamic time warping, DTW)[10-11]. 动态时间弯曲作为一种鲁棒性强的度量方法,不仅能够度量不等长度的时间序列,还能够通过弯曲时间轴来匹配时间序列数据,很好地反映了时间序列的形态特征.

本文针对以上问题及动态时间弯曲度量的优越性,提出一种关键形态特征的多元时间序列降维方法. 利用动态时间弯曲找出训练集中每个类别的中心多元时间序列,根据中心多元时间序列的形态特征选择若干个具有关键特征的变量分量,以各个类别的关键特征变量分量作为该类别数据降维后的变量分量,进而达到数据降维的目的. 新方法不仅能够有效地进行数据降维,还能够得到较好的数据分类质量.

1 相关定义及方法

1.1 基本定义

定义1 多元时间序列(multivariate time series, MTS). 由多个存在相互作用或一定相关关系的一元时间序列组成的时间序列. 给定一个多元时间序列$Q=(f_1,f_2,\dots,f_m)$且$f_i=(q_{1i},q_{2i},\dots,q_{ni})^{\mathrm{T}}$,$m$表示变量维度,$n$表示时间维度,即$t=(1,2,\dots,n)$,则$f_i$表示变量维度为$i$的一系列观测值. 多元时间序列可以通过一个$n\times m$的矩阵来表示,即

$$Q_{n\times m}=\begin{bmatrix} q_{11} & q_{21} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nm} \end{bmatrix}.$$

定义2 多元时间序列数据集. 给定一个多元时间序列数据集$D=\{Q_1,Q_2,\dots,Q_K\}$,K为MTS个数,令f_j^i表示Q_i的第j个变量分量.

值得注意的是,一个MTS也可看作是一元时间序列数据集,约定符号f_j^i为一元时间序列,当用在多元时间序列中,则表示为MTSQ_i的j个变量分量;用在一元时间序列数据集中,则表示一元时间序列数据集I的第j条时间序列,而此时f的上标则标识数据集的名称.

1.2 动态时间弯曲(DTW)

动态时间弯曲是一种度量精度高、鲁棒性强的度量方法[12],在语音识别[13]、图像挖掘[14-15]、金融分析[16-17]等领域中得到了广泛的应用. 动态时间弯曲不仅能通过弯曲时间轴来度量不等长的时间序列,还能够充分反映时间序列的形态特征.

给定两条一元时间序列$f_a=(a_1,a_2,\dots,a_N)$和$f_b=(b_1,b_2,\dots,b_M)$,构建$N\times M$的距离矩阵

$$D_{N\times M}=d(a_i,b_j).$$

其中:$d(a_i,b_j)=\sqrt{(a_i-b_j)^2}$,$i=1,2,\dots,l$, $j=1,2,\dots,h$. DTW的目的在于寻找一条f_a和f_b之间的最短弯曲路径$P=(p_1,p_2,\dots,p_K)(\max\{N,M\}\leqslant K\leqslant N+M-1)$,$p_k=(i,j)$,使得累积距离$\gamma(N,M)$最小,且这条最短弯曲路径满足边界性、连续性、单调性. 最终,有

$$\mathrm{DTW}(f_a,f_b)=\min\Big(\sum_{k=1}^{K}p_k\Big)=\gamma(N,M), \tag{1}$$

而累积距离通过

$$\gamma(i,j)=d(a_i,b_j)+\min\begin{cases}\gamma(i-1,j-1),\\ \gamma(i-1,j),\\ \gamma(i,j-1)\end{cases} \tag{2}$$

得到.

1.3 DTW barycenter averaging(DBA)

时间序列聚类需要求得合适的簇中心,而DTW质平均(DTW barycenter averaging, DBA)[18]方法能够使得在聚类过程中得到合理的簇中心. 该方法旨在从数据点匹配的角度来综合考量簇中时间序列的“平均”状态,利用这种状态下的时间序列来表征中心序列. 以DTW为基础技术,设一元时间序列f_a为初始簇中心,计算$f_a=(a_1,a_2,\dots,a_N)$与一元时间序列数据集$B=\{f_1^b,f_2^b,\dots,f_L^b\}$中所有时间序列的DTW距离,并分别记录$f_a$与$B$中所有时间序列之间的数据匹配关系$\mathrm{path}(\cdot)$,根据匹配关系来求得中心序列,其中$L$表示数据集$B$的时间序列数量,$\mathrm{path}(\cdot)$表示与$f_a$中数据点匹配的点的集合. 通过公式

$$c_i=\sum_{j=1}^{L}\sum_{b_k\in\mathrm{path}_j(a_i)}b_k\Big/\sum_{j=1}^{L}|\mathrm{path}_j(a_i)| \tag{3}$$

求得中心序列的数据点,其中$|\mathrm{path}_j|$表示path_j的模,即path_j中数值的个数. c_i的实质是指与a_i数据点具有匹配关系的所有数据点的平均值. 以下给出DBA的算法步骤.

算法1 $\mathrm{DBA}(f_a,B)$.

输入: 初始中心序列$f_c=(c_1,c_2,\dots,c_N)$,数据

集$B=\{f_1^b,f_2^b,\ldots,f_L^b\}$;

输出: 中心序列$f_c=(c_1,c_2,\ldots,c_N)$.

step 1: 第l次循环,计算DTW(f_c,f_l^b),合并第1次至第$l-1$次循环时数据点c_i匹配的所有数据点,即$\mathrm{path}_l(c_i)\leftarrow\mathrm{path}_l(c_i)\bigcup\mathrm{path}_{l-1}(c_i),i=1,2,\ldots,N$.

step 2: 重复Step 1,当$l=L$时, 利用式(3)求解中心序列f_c的每个值,算法结束.

step 1遍历数据集B中的时间序列,计算B中所有时间序列与初始中心序列的DTW距离. 在每次的遍历计算中,均记录初始中心序列每个数据点与B中时间序列匹配的数据点,取并集. 例如,f_c与L条时间序列进行度量之后,数据点c_1产生了L个$\mathrm{path}(c_1)$集合,对这L个$\mathrm{path}(c_1)$集合取并集,最终得到c_1匹配的数据点集合$\mathrm{path}_{\mathrm{Union}}(c_1)$.

step 2的目的是更新中心序列,当初始中心序列的每个数据点都得到了相应的匹配数据集合之后,通过式(3)更新中心序列的值.

上述算法过程描述中仅描述了对初始中心序列的一次迭代更新,因此需要多次迭代更新中心序列,收敛后得到最终的中心序列.

2　特征表示方法

传统的降维方法主要通过转换数据的坐标轴,利用数据的方差来反映信息分布情况,选择能够尽可能多的反映信息的成分作为新的变量. 然而,在某些情况下,变量分量的形态特征与其他变量分量的差别很大,仅利用几个变量分量的形态特征即可相互区分MTS的类别. 若提取到这种特殊的变量分量,不仅能够达到MTS降维的目的,还能够体现MTS的特征信息. 因此,有必要从形态特征角度出发来研究MTS的特征表示方法.

2.1　关键形态特征表示

同一数据集的MTS具有一致的结构特征,即变量维数相同. 类别相同的MTS对应的变量分量展现的形态特征是相似的,类别相异的则有若干个变量分量的形态特征是不同的. 衡量两个MTS对应变量分量的相似程度,需要计算这两个变量分量距离的大小;判断两个MTS是否属于同一类别,需要得到所有变量分量的距离之和. 当变量数目很大,变量分量的长度很长甚至MTS之间的长度不等时,直接度量MTS的距离来判断类别的工作量是很大的. 然而,在某些情况下,在同一数据集中不同类别的MTS,存在若干相同或不同的特殊变量,其分量的形态特征差异性很大,利用这些分量的形态特征可以直接相互区分其类别. 这种区分性强的形态特征,称之为关键形态特征(crucial shape feature),对应的变量称之为关键特征变量,其分量为关键特征分量,如图1所示.

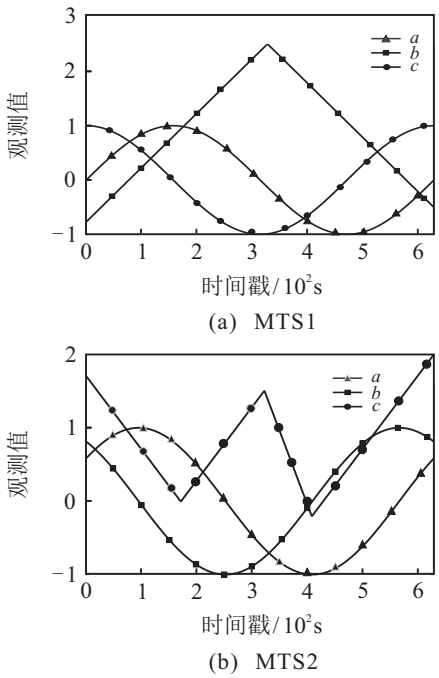

图1　多元时间序列及其关键形态特征

图1显示的是两个不同类别的MTS,每个MTS具有3个变量. 通过观察可知,MTS1与MTS2对应变量a、b和a、c分量的形态特征都非常相似. 然而,MTS1的变量b分量和MTS2的变量c分量的形态特征差别很大,甚至仅仅分别利用变量b和变量c分量即可区别这两个MTS. 因此,分别称变量b、c分量的形态特征为MTS1和MTS2的关键形态特征. 可知不同类别MTS的关键变量可以相同,也可以不同. 一个MTS的关键形态特征是该MTS的一个或一个以上原始变量分量构成的矩阵,既充分利用了原始数据信息,又达到了MTS数据降维的目的.

提取一个MTS的关键形态特征,需要综合利用同一类别的特征信息以及不同类别MTS的变量分量之间的形态差异. 仍以图1为例,为了提取两个类别的MTS的关键特征变量,首先分别综合考虑每个类别的MTS整体形态,每个类别用一个新的多元时间序列来反映其综合形态;其次利用每个类别的综合形态来得到类别间各个变量之间的形态差异关系;最后综合利用类内与类间的变量形态特征信息来得到关键形态变量.

为此,本文提出一种关键形态特征表示方法(crucial shape feature, CSF). 给定一个变量为m的MTS数据集$D=\{Q_1,Q_2,\ldots,Q_K\}$,K为MTS个数. 值得注意的是,即使在同一个数据集中,MTS的长度也不一定等长,即Q_i和Q_j时间维度$n_i=n_j$或

者$n_i \neq n_j$. 由于MTS各个变量分量之间存在着量纲,在数据挖掘工作展开之前,需要先进行数据标准化处理, 以消除量纲的影响, 保证变量分量缩放和偏移的不变性[19]. 分别对Q_i的m个变量分量进行标准化,使各个分量分别服从标准正态分布.

CSF将类别数为C的D按类进行划分,得到C个子集,即$\underset{c}{D} = \{\underset{c}{Q_i}\}(i = 1, 2, \ldots, N_c, c \in C)$,表示类别为$c$的MTS组成的子集,$N_c$表示$|\underset{c}{D}|$. 利用DBA求出各个子集的中心多元时间序列(central multivariate time series, CMTS),再分别找出CMTS的r个关键变量,通常$r < m$. 将D中每个MTS按类别提取对应的关键变量分量,最终得到变量维度为r的$D' = \{Q'_1, Q'_2, \ldots, Q'_K\}$. 关键形态特征的具体算法如下.

算法2 CSF(D).

输入: MTS训练数据集$D = \{Q_1, Q_2, \ldots, Q_K\}$,其变量维度为$m$;

输出: 新MTS训练集$D' = \{Q'_1, Q'_2, \ldots, Q'_K\}$,其变量维度为$r$.

step 1: 按类别划分D,得到$\underset{c}{D} = \{\underset{c}{Q_i}\},\ i = 1, 2, \ldots, N_c,\ c \in C$.

step 2: 从$\underset{c}{D} = \{\underset{c}{Q_i}\}$中随机选择一个MTS作为初始CMTS,记为第$t$个$\underset{c}{Q_t}$,计算各个类别CMTS,这一步骤又包含以下两个小步骤:

step 2.1: 针对$\underset{c}{D}$中不包含$\underset{c}{Q_t}$的MTS,利用这些MTS的第$d\,(d = 1, 2, \ldots, m)$个变量分量构建一个新的临时一元时间序列数据集$\underset{c}{B} \leftarrow \{f_d^b\}$,计算类别$c$的第$d$个变量的中心序列$\underset{c}{f_d} \leftarrow \mathrm{DBA}(f_d^t, \underset{c}{B})$.

step 2.2: 重复step 2.1,遍历完每个变量时$(d = m)$结束,最终得到类别为c的CMTS,此处视CMTS为一元时间序列数据集$\underset{c}{C}$,即$\underset{c}{C} = \{f_1^c, f_2^c, \ldots, f_m^c\}$.

step 3: 重复step 2,直到$c = C$,得到各个类别的CMTS.

step 4: 计算

$$\underset{c}{f_F} \leftarrow \{\underset{c}{v_i}\}.$$

其中: $\underset{c}{v_i} \leftarrow \sum_{c_1 \neq c}^{C} \sum_{j=1}^{m} \mathrm{DTW}(f_i^c, f_j^{c_1}), i = 1, 2, \ldots, m$, f_i^c表示数据集C第i条时间序列, $f_j^{c_1}$表示数据集C_1的第j条时间序列(这里以类别来标识数据集名称); $\underset{c}{v_i}$表示类c的CMTS的第i维与其他类别CMTS的所有维度的距离之和; $\underset{c}{f_F}$表示类c的MTS变量权重向量. 重复该步骤,直到获得所有类别的变量权重向量.

step 5: 将$\underset{c}{f_F}$中的值按从大到小排序,选出前r个值,该r个值对应的变量为类c的关键特征变量,$\underset{c}{D} = \{\underset{c}{Q_i}\}(i = 1, 2, \ldots, N_c, c \in C)$按照关键特征变量提取分量,得到$\underset{c}{D'} = \{\underset{c}{Q'_i}\}$. 重复该步骤至$c = C$,直到$C$个类别的子集均提取关键特征分量. 合并新子集,得到$D' = \{Q'_1, Q'_2, \ldots, Q'_K\}$,算法结束.

算法CSF$(\cdot)$描述的是从训练集中提取各个类别的关键特征变量分量,每个类别的MTS只保留关键特征分量,将训练集数据转化为只包含关键特征变量分量的数据,从而实现MTS降维的目的. 关键形态特征的提取需要充分利用同一类别的MTS各分量之间的信息,因此需要充分考虑数据集的类别信息. step 1的目的是实现数据集按类别的划分,为后续工作做好准备. 为了综合考虑子集中所有MTS的各个分量的形态特征,每个类别使用CMTS来代表该类别子集的整体形态. step 2和step 3是在同一类别子集中,利用DBA$(\cdot)$求得每个变量维度的中心序列来构建该类别的CMTS,最终得到每个类别的CMTS,充分利用了类内的特征信息. step 4是为了得到所有类别的各个变量的特征重要度,从而提取关键的变量. 特征重要度越大,表示该维度与其他类别的所有维度的形态差距越大. step 5的目的在于将数据按类别来提取其关键特征变量分量,得到降维后的数据集. 算法的性能需要通过测试集进行验证,在进行时间序列相似性度量时,传统主成分分析方法得到主成分之后,仍然需要对每个测试数据进行转换,而利用CSF$(\cdot)$得到训练集各个类别的关键特征分量之后,测试数据只需提取与训练数据相同的关键特征分量即可.

2.2 时间复杂度

训练集

$$D = \{Q_1, Q_2, \ldots, Q_K\},\ \underset{c}{D} = \{\underset{c}{Q_i}\}.$$

其中: $i = 1, 2, \ldots, N_c, c = 1, 2, \ldots, C$,变量维度为$m$,时间序列长度为$n$. 对于一个MTS而言, DTW$(\cdot)$的时间复杂度为$O(mn^2)$,可知求解所有类别的CMTS时, DBA$(\cdot)$的时间复杂度为$O((K - C)mn^2)$,计算类$c$的MTS各个变量的特征权重向量的时间复杂度为$O(Cm^2n^2)$. 因此, CSF$(\cdot)$的时间复杂度为$O((K - C)mn^2 + Cm^2n^2)$. 可以发现,CSF方法的时间复杂度与变量数、时间序列长度、类别数、每种类别的数据个数有关. 随着维度以及时间序列长度的增加, CSF$(\cdot)$的时间消耗也越来越多. 然而,训练过程中一旦提取到关键特征变量,测试过程只需做简单的特征提取工作,因此CSF可以为后续挖掘工作提供便利.

3 数值实验

为了验证CSF方法的有效性和优越性,本文使用9个多元时间序列数据集和32支股票数据集分别进

行两组分类实验. 分别与文献[2]的共同主成分分析(CPCA)方法和文献[3]的相关特征表示(RFR)方法以及主成分分析(PCA)方法的分类结果进行比较和分析,以体现CSF的优越性.

3.1　数据选取

第1组分类实验使用9个不同领域的多元时间序列数据集,具体情况如表1所示. 由于MTS长度不等,表格不列举MTS长度. 为了消除量纲对实验的影响,均先对数据进行标准化处理,使得每个变量分量均值为0,方差为1.

表1　MTS数据集信息

数据集	训练数	测试数	类别数	变量数
ECG	100	100	2	2
Japanese Voewls	270	370	9	12
Lp1	38	50	4	6
Lp2	17	30	5	6
Lp3	17	30	4	6
Lp4	42	75	3	6
Lp5	64	100	5	6
Libros	180	180	15	2
Pen Digits	300	10692	13	2

第2组分类实验使用真实的股票多元时间序列数据,随机抽取2015年1月5日至2015年12月31日共32只股票的多元时间序列数据,分别有房地产、金融、工业3个行业;包含4个变量维度,即开盘价、最高价、最低价和收盘价;时间序列长度在177～244之间. 数据分别为10个房地产行业的MTS,14个金融行业的MTS和8个工业行业的MTS. 该数据从国泰安CSMAR精准财经研究数据库下载,具体信息如表2所示. 为消除量纲的影响,对数据进行标准化处理. 从每个行业中随机抽取若干个数据作为训练集,其中带“*”号的数据为被抽取的数据. 因此,股票数据训练集个数为11个,测试集为32个.

表2　3种类别股票的数据集信息

股票代码	行业	长度	股票代码	行业	长度
000001*	金融	244	000537	地产	239
000002	地产	235	000559	工业	244
000004*	地产	244	000718	地产	231
000012*	地产	220	000776*	金融	244
000017*	地产	244	002142	金融	244
000020*	地产	179	002736*	金融	237
000027	工业	244	600015*	金融	244
000031	地产	244	600999	金融	238
000042	地产	239	601009	金融	239
000048	地产	177	601099	金融	244
000049	工业	243	601169	金融	213
000060*	工业	239	601628	金融	244
000100*	工业	244	601688	金融	244
000338	工业	244	601818	金融	234
000423*	工业	244	601939	金融	244
000528	工业	244	601998	金融	242

3.2　分类实验

本次实验总共分为两组分类实验,第1组分类实验使用9个不同领域的MTS数据集,第2组分类实验使用股票数据集. 实验采用最近邻分类方法,对比方法采用CPCA、RFR和PCA. 数据集中的时间序列长度不等,因此CPCA和PCA所得到的成分序列长度也不一样,需要利用DTW来计算特征序列之间的相似度.

第1组实验结果如表3所示,展现了4种方法下的两种错误率,即平均分类错误率及最小分类错误率. 平均分类错误率为各种降维数下分类错误率的平均值,最小错误率为各种降维数下分类错误的最小值,r值为最小错误率下的变量数. 实验对胜出率和总平均值进行了简单的汇总,胜出率为错误率优胜的个数与数据集个数的比值,总平均值为对各列的错误率再求平均. 胜出率越大越好,表明胜出个数越多;总平均值越小越好,表明错误率在整体上优胜.

表3　分类错误率　　单位: %

	CSF		CPCA		RFR		PCA	
	平均值	最小值(r)	平均值	最小值(r)	平均值	最小值(r)	平均值	最小值(r)
1	0.195 0	0.190 0(1)	0.560 0	0.520 0(2)	0.670 0	0.670 0(1)	0.595 0	0.590 0(2)
2	0.834 3	0.797 3(9)	0.842 9	0.778 4(2)	0.913 5	0.889 2(1)	0.838 4	0.797 3(2)
3	0.308 0	0.220 0(5)	0.368 0	0.320 0(5)	0.740 0	0.580 0(1)	0.300 0	0.240 0(5)
4	0.380 0	0.300 0(5)	0.413 3	0.300 0(2)	0.533 3	0.500 0(2)	0.586 7	0.500 0(2)
5	0.493 3	0.466 7(1)	0.513 3	0.466 7(1)	0.560 0	0.533 3(1)	0.446 7	0.400 0(1)
6	0.354 7	0.293 3(3)	0.290 7	0.266 7(2)	0.736 0	0.480 0(1)	0.421 3	0.373 3(2)
7	0.500 0	0.470 0(1)	0.550 0	0.490 0(5)	0.776 0	0.770 0(2)	0.526 0	0.500 0(5)
8	0.319 4	0.211 1(2)	0.308 4	0.216 7(2)	0.933 3	0.933 3(1)	0.463 9	0.416 7(2)
9	0.159 3	0.083 1(2)	0.146 7	0.078 1(2)	0.900 1	0.896 0(2)	0.232 5	0.130 5(2)
平均值	**0.393 8**	**0.336 8**	0.443 7	0.388 1	0.751 4	0.694 6	0.490 1	0.438 6
胜出率	**0.444 4**	**0.555 6**	0.333 3	0.444 4	0.000 0	0.000 0	0.222 2	0.111 1

从表3可知,CSF的平均错误率和最小错误率大多数情况下低于其余3种方法. RFR的平均错误率和最小错误率均比CSF的高,说明仅利用变量的相关性来进行多元时间序列的相似性度量比直接使用原始数据来度量的效果差. CSF方法的平均错误率整体上比CPCA的低,而且CSF在训练样本较少的情况下仍能得到很好的分类效果,并且一旦得到关键特征分量,测试集只需要进行简单的分量提取工作.

利用CSF方法进行分类时,关键特征变量个数r的不同,对分类的结果会造成一定的影响. 选择的数量太少,则使得分类过程中信息不足;反之,又会对分类过程中形成信息干扰. 当达到合适的个数时,关键信息足够且冗余信息量少,分类结果最好,如图2所示.

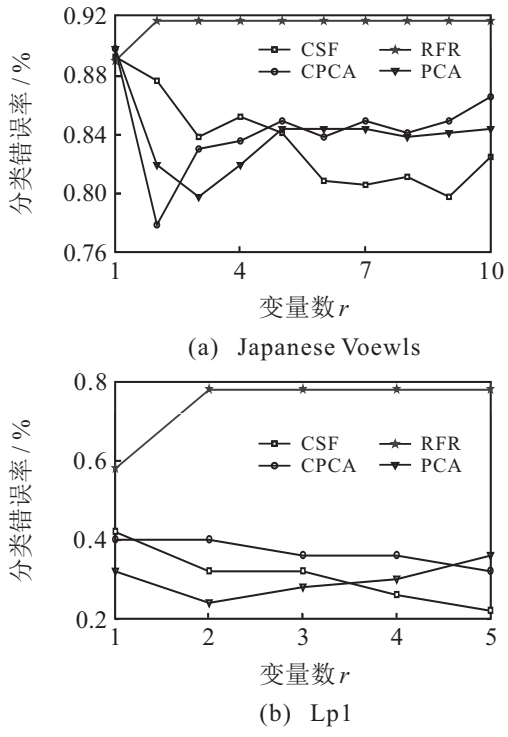

图2 关键特征变量数r对分类错误率的影响

通过数据集Japanese Voewls和Lp1来分析关键特征变量数对分类错误率的影响,图2显示了4种方法在降维后,变量数对分类错误率的影响. 随着关键特征变量的增加,数据集的分类错误率有一定的下降,表明随着数据信息的增多,对分类错误率的降低起到一定的帮助作用. 然而,当到达一定程度后,冗余信息开始增多,反而造成了干扰,如图2(a)所示. 尽管如此,实际情况中也会存在某种数据集,这类数据集的各个变量表达着各不相同的信息,保持尽可能多的变量对分类会更有帮助,如图2(b)所示. 图3更好地说明了这种情况. 图3是在含有4种类别的数据集Lp1中,从每种类别随机抽取1个MTS,并对每个变量分量进行标准化、平移后作出的图,每个MTS有6个变量,每个变量分量用不同图标来标注. 仔细观察可以发现,不同类别MTS的相同变量分量所呈现的形态有一定的区别. 实际情况中,若使用全部变量则没有起到降维作用,因此在分类错误率可接受的范围内,选取适当的变量数来实现降维.

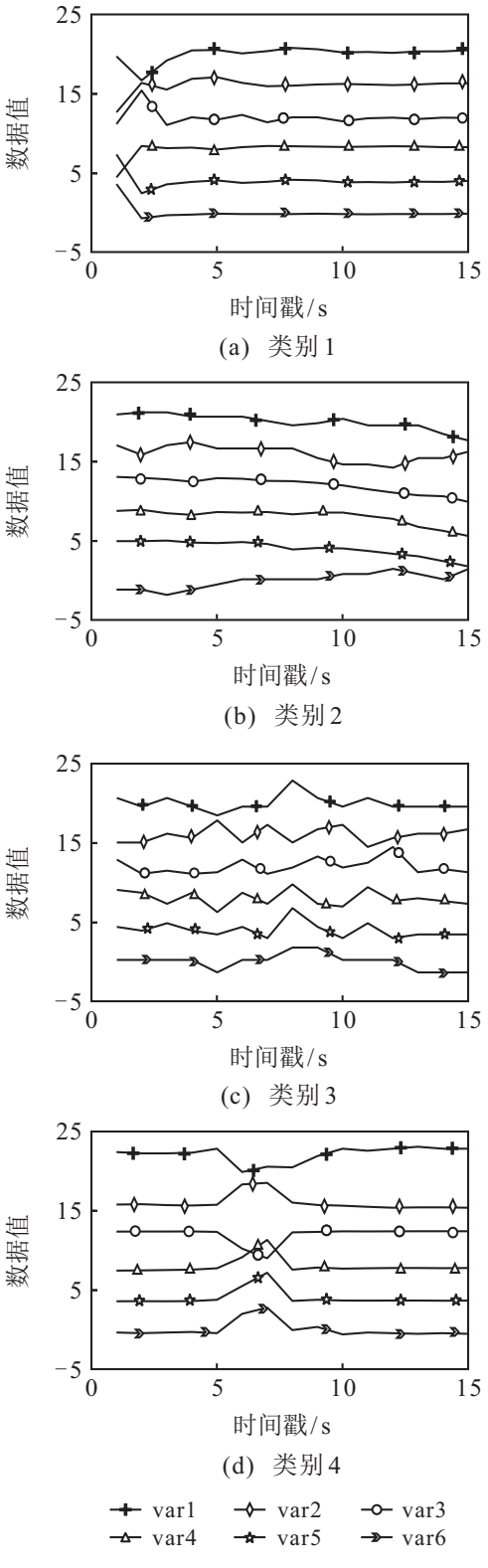

图3 数据集Lp1四种类别的变量分量形态

第2组分类实验采用真实股票数据,观察4种方法下分类错误率情况. 为了达到降维目的,统一降低2个维度,即降维后变量数为2,实验结果如图4所示.

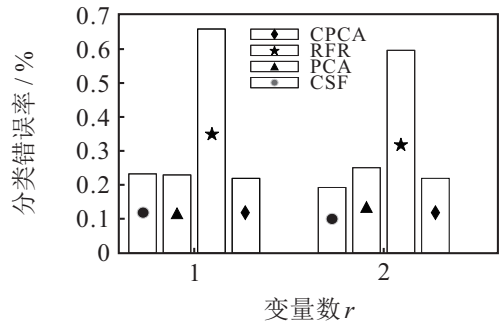

图 4　股票多元时间序列分类错误率

如图4所示,CSF方法在关键特征变量数为2时的股票分类错误率要低于关键特征变量数为1的错误率,说明这4个变量维度中,至少利用一半的信息能取得好的效果. 当关键特征变量数为2时,4种方法中CSF取得了最好的效果. 说明在真实股票市场环境中,CSF方法具有一定的可行性和优势.

3.3　时间效率分析

时间效率是衡量多元时间序列数据挖掘方法性能的重要指标之一. 利用4种方法对Lp3数据集进行测试阶段的特征表示及分类时间消耗总代价进行比较,考察在不同变量维度下总时间消耗区别. 4种不同方法在不同变量数的测试总时间代价结果数据如表4和图5所示.

表4　Lp3的平均分类效率　　单位: s

方法	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
CSF	0.016 9	0.033 5	0.050 1	0.066 5	0.083 3
CPCA	0.017 8	0.033 4	0.050 2	0.067 4	0.084 1
RFR	0.002 1	0.003 5	0.004 4	0.005 1	0.006 2
PCA	0.018 2	0.034 6	0.051 6	0.068 3	0.084 4

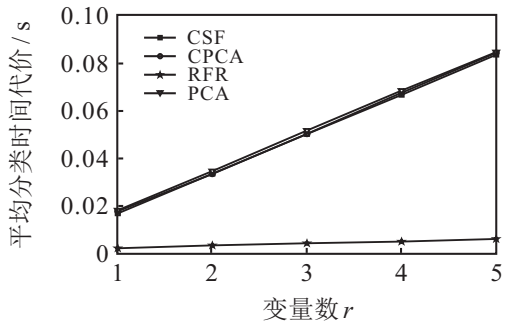

图 5　平均分类效率比较分析

图5是表4的直观展示,从图5可以发现,RFR的时间代价是最小的,原因在于RFR将不等长的MTS转换成等长度的相关性特征矩阵,能够直接使用欧氏距离来度量. PCA仍需要对每个测试数据的协方差进行奇异值分解,而CPCA只需要进行一次奇异值分解,因此PCA的时间代价高于CPCA. 由于CSF前期训练阶段需要计算CMTS和关键特征权重,该阶段的时间代价不占优势. 然而,当得到关键特征变量之后,测试数据只需要提取分量,其时间代价与CPCA计算协方差时间代价一样. 而CPCA比CSF多出一部分时间来对平均协方差矩阵进行奇异值分解,该部分的时间消耗为变量数的平方. 在测试阶段,CSF的时间代价要略低于CPCA方法,在降维数较高的情况下,CSF更能体现优越性.

4　结　论

本文针对多元时间序列各分量的形态特征的重要性,提出了关键形态特征的多元时间序列降维方法. 与传统方法相比,新方法具有以下优势: 1)通过求解中心序列,有效地综合考虑了数据集同一类别和不同类别的数据整体形态信息; 2)通过利用中心序列来求解关键特征变量的重要度,充分反映了多元时间序列变量的形态表征重要度的先后顺序,使得利用重要度大的变量分量来进行数据特征表示,从而体现了多元时间序列的变量分量形态的差异性; 3)相比之下,在训练样本集较少的情况下能够得到较好的分类效果,而且在降维变量数较高的情况下,能够体现CSF一定的数据挖掘效率.

参考文献(References)

[1] Yang D, Dong Z, Lim L H, et al. Analyzing big time series data in solar engineering using features and PCA[J]. Solar Energy, 2017, 153(1): 317-328.

[2] 李正欣, 郭建胜, 惠晓滨, 等. 基于共同主成分的多元时间序列降维方法[J]. 控制与决策, 2013, 28(4): 531-536.
(Li Z X, Guo J S, Hui X B, et al. Dimension reduction method for multivariate time series based on common principal component[J]. Control and Decision, 2013, 28(4): 531-536.)

[3] 李海林. 基于变量相关性的多元时间序列特征表示[J]. 控制与决策, 2015, 30(3): 441-447.
(Li H L. Feature representation of multivariate time series based on correlation among variables[J]. Control and Decision, 2015, 30(3): 441-447.)

[4] Spiegel S, Gaebler J, Lommatzsch A, et al. Pattern recognition and classification for multivariate time series[C]. Proceedings of the 5th International Workshop on Knowledge Discovery from Sensor Data. San Diego: ACM, 2011: 34-42.

[5] 吴虎胜, 张凤鸣, 钟斌. 基于二维奇异值分解的多元时间序列相似匹配方法[J]. 电子与信息学报, 2014, 36(4): 847-854.
(Wu H S, Zhang F M, Zhong B. Slimilar pattern matching method for multivariate time series based on two-dimensional singular value decomposition[J]. Journal of Electronics and Information Technology, 2014, 36(4): 847-854.)

[6] Liu C, JaJa J, Pessoa L. LEICA: Laplacian eigenmaps for group ICA decomposition of fMRI data[J]. NeuroImage, 2018, 169(4): 363-373.

[7] Sáfadi T. Using independent component for clustering of time series data[J]. Applied Mathematics and Computation, 2014, 243(15): 522-527.

[8] 李海林, 梁叶, 王少春. 时间序列数据挖掘中的动态时间弯曲研究综述[J]. 控制与决策, 2018, 33(8): 1345-1353.
(Li H L, Liang Y, Wang S C. A review on dynamic time warping in time series data mining[J]. Control and Decision, 2018, 33(8): 1345-1353.)

[9] 李海林, 梁叶. 基于数值符号和形态特征的时间序列相似性度量方法[J]. 控制与决策, 2017, 32(3): 451-458.
(Li H L, Liang Y. Similarity measure based on numerical symbolic and shape feature for time series[J]. Control and Decision, 2017, 32(3): 451-458.)

[10] Zhao J, Itti L. ShapeDTW: Shape dynamic time warping[J]. Pattern Recognition, 2018, 74(2): 171-184.

[11] Wan Y, Chen X L, Shi Y. Adaptive cost dynamic time warping distance in time series analysis for classification[J]. Journal of Computational and Applied Mathematics, 2017, 319(1): 514-520.

[12] 李正欣, 张凤鸣, 李克武, 等. 一种支持DTW距离的多元时间序列索引结构[J]. 软件学报, 2014, 25(3): 560-575.
(Li Z X, Zhang F M, Li K W, et al. Index structure for multivariate time series under DTW distance metric[J]. Journal of Software, 2014, 25(3): 560-575.)

[13] Yeh P H, Yang S L, Yang C C, et al. Automatic recognition of repetitions in stuttered speech: Using end-point detection and dynamic time warping[J]. Procedia-Social and Behavioral Sciences, 2015, 193(C): 356.

[14] Cheng H, Dai Z, Liu Z, et al. An image-to-class dynamic time warping approach for both 3D static and trajectory hand gesture recognition[J]. Pattern Recognition, 2016, 55: 137-147.

[15] Adwan S, Alsaleh I, Majed R. A new approach for image stitching technique using dynamic time warping (DTW) algorithm towards scoliosis X-ray diagnosis[J]. Measurement, 2016, 84: 32-46.

[16] Tsinaslanidis P E. Subsequence dynamic time warping for charting: Bullish and bearish class predictions for NYSE stocks[J]. Expert Systems with Applications, 2018, 94(3): 193-204.

[17] Gong X, Si Y W, Fong S, et al. Financial time series pattern matching with extended UCR suite and support vector machine[J]. Expert Systems with Applications, 2016, 55(15): 284-296.

[18] Petitjean F, Ketterlin A, Gançarski P. A global averaging method for dynamic time warping, with applications to clustering[J]. Pattern Recognition, 2011, 44(3): 678-693.

[19] 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.)

作者简介

李海林 (1982–), 男, 教授, 博士, 从事数据挖掘与决策支持等研究, E-mail: hailin@mail.dlut.edu.cn;

梁叶 (1992–), 女, 硕士生, 从事数据挖掘与金融数据的研究, E-mail: allin0258@163.com.

(责任编辑：齐 霁)