文章编号: 1001-0920 (2015) 03-0441-07 DOI: 10.13195/j.kzyjc.2014.0132

基于变量相关性的多元时间序列特征表示

李海林

(华侨大学 工商管理学院，福建 泉州 362021)

摘 要: 针对高维特性对多元时间序列数据挖掘过程和结果的影响, 以及传统主成分分析方法在多元时间序列数据特征表示上的局限性, 提出一种基于变量相关性的多元时间序列数据特征表示方法. 通过协方差矩阵描述每个多元时间序列的分布特征和变量相关关系, 利用主成分分析方法对综合协方差矩阵进行主元分析, 进而实现多元时间序列的数据降维和特征表示. 实验结果表明, 所提出的方法不仅能提高多元时间序列数据挖掘的质量, 还可以对不等长多元时间序列进行快速有效的挖掘.

关键词: 多元时间序列；主成分分析；特征表示；数据挖掘

中图分类号: TP273 **文献标志码:** A

Feature representation of multivariate time series based on correlation among variables

LI Hai-lin

(School of Business Management，Huaqiao University，Quanzhou 362021，China．E-mail：hailin@mail.dlut.edu.cn)

Abstract: The property of high dimensionality impacts on the process and results in the field of time series data mining, and the traditional methods about principal component analysis have some limitations to represent multivariate time series. Therefore, a feature representation of multivariate time series based on correlation among variables is proposed. The distribution and relationships among variants of every time series are described by the covariance matrix, and principal components are extracted from an integrated covariance matrix by principal component analysis. In this way, the dimensionality of multivariate time series can be reduced and the features can be represented. The experimental results show that the proposed method not only improves the quality of multivariate time series data mining but also efficiently mines on the data with different lengths.

Keywords: multivariate time series；principal component analysis；feature representation；data mining

0 引 言

多元时间序列是数据挖掘领域中重要的数据类型之一[1], 广泛存在于金融、医疗、电子信息和气象等科学工程领域. 从狭义上讲, 多元(多变量)时间序列是由多个一元(单变量)时间序列组合而成的, 各一元时间序列相互作用或存在一定的相互关系[2]; 从广义上讲, 多元时间序列是某一特定系统根据时间先后顺序产生的数据序列, 系统中的各个因素产生相应的一元时间序列, 如地理信息系统、智能监控系统和航空发动机诊断系统等都产生大量的多元时间序列数据. 然而, 其时间维度和变量维度的高维性决定了整个数据挖掘过程的复杂性, 并影响最终挖掘结果的准确性.

时间序列数据挖掘通常包括聚类、分类、关联规则、兴趣模式发现、异常检测、相似性搜索和可视化等, 其挖掘效率和质量容易受到时间序列数据特征复杂性的影响. 为了提高时间序列数据挖掘技术的性能, 通常利用数据降维和特征表示降低数据挖掘过程或模型的复杂性, 并通过清除冗余信息的影响来提高挖掘结果的准确性. 目前, 人们已提出不少相关降维技术和特征表示方法. 例如, 单变量时间序列数据降维[3-4]、主成分分析(PCA)[5-6]、奇异值分解(SVD)[7-8]和独立成分分析(ICA)[9]等方法. 其中, 主成

收稿日期: 2014-01-21；**修回日期:** 2014-03-28.

基金项目: 国家自然科学基金项目(61300139)；福建省中青年教师教育科研项目(JAS14024)；华侨大学中青年教师科研提升计划项目(ZQN-PY220).

作者简介: 李海林(1982–), 男, 讲师, 博士, 从事数据挖掘与决策支持的研究.

分分析方法是多元时间序列数据挖掘中最重要的数据降维方法, 它通过数据坐标变换得到更能反映多元时间序列数据特征分布的各个主成分, 进而实现降维和特征表示. 同样, 基于共同主成分分析多元时间序列降维方法也具有良好的降维效果[10]. 然而, 对于不等长的多元时间序列数据挖掘, 传统方法虽然能够对其进行数据降维, 但降维后的数据依然存在长度不相等的问题, 这将给后期的数据挖掘过程带来一定的麻烦. 例如, 需要利用时间复杂度较高的动态时间弯曲 (DTW)[11]来度量不等长特征序列之间的相似性[10], 并且传统方法的特征表示效果易于受原始异常数据的影响.

针对上述问题, 本文提出一种基于变量相关性的多元时间序列特征表示方法. 通过对每个多元时间序列建立协方差矩阵来描述其变量相关性[12], 并构建相应的综合协方差矩阵代替原始时间序列数据集, 从而更好地反映多元时间序列变量之间的相互关系. 同时, 利用主成分分析方法对多元时间序列数据集的变量相关性矩阵进行主元分析, 选择信息量较大的特征向量作为新特征空间的坐标系数, 进而实现原始多元时间序列的特征表示. 该方法能够解决不等长时间序列特征表示带来的相似性度量问题, 不仅能够提高数据挖掘算法的效率, 还能提升挖掘结果的质量.

1 主成分分析

主成分分析 (PCA) 是一种重要的数据降维方法, 在信号处理[13]、质量检测[14]和语音识别[15]等领域中得到广泛的应用. 同样, 在多元时间序列数据挖掘中, 主成分分析能够通过数据变换较好地对多元时间序列进行数据降维和特征表示[5-6,10]. 根据多元时间序列数据的分布特征, PCA 可以在保留所有信息或大部分信息的前提下, 利用前几个主成分对原始多元时间序列进行特征表示, 进而实现数据降维. 其原理是将原来维度较高且彼此可能存在信息冗余或相关的变量用线性组合的方法形成维度较低且不相关的特征变量, 即

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{nj}X_n. \tag{1}$$

其中: X_i 为多元时间序列中的第 i 个变量序列, Y_j 为通过 PCA 变换后得到的不相关的主成分特征变量, a_{ij} 为成分权重.

对于某一多元时间序列 $X_t(m) = (X_1, X_2, \cdots, X_m)$. 其中: m 为变量维度; n 为时间维度, 即 $t = 1, 2, \cdots, n$, 且在通常情况下 $n \gg m$; $X_i = (x_{1i}, x_{2i}, \cdots, x_{ni})^{\mathrm{T}}$. 多元时间序列数据可表示为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}.$$

PCA 通过正交化线性变换, 将原始多元时间序列的观察数据变换到另外一个新的坐标系中, 使得这一组观测数据的任何投影的最大方差在第一个主成分上, 第二大方差在第二个主成分上, 依此类推, 可以得到各主成分 $Y = (Y_1, Y_2, \cdots, Y_m)$. 算法过程如下.

主成分分析算法 $Y = \mathrm{PCA}(X)$.

输入: 多元时间序列 $X = (X_1, X_2, \cdots, X_m)$;

输出: 主成分特征序列 $Y = (Y_1, Y_2, \cdots, Y_m)$.

Step 1: 对多元时间序列的各个变量进行标准化处理, 进而消除量纲影响, 即 $X = \mathrm{zscore}(X)$.

Step 2: 变量间的相似性计算. 通过协方差来描述变量间的相关关系, 即 $S = \mathrm{E}[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))]$. 由 Step 1 可知, $\mathrm{E}(X_i) = 0, \mathrm{E}(X_j) = 0$, 故 $S = \mathrm{E}[X_i X_j]$.

Step 3: 通过 SVD 对协方差矩阵 S 求特征值 W 和特征向量 U, 有 $S = U\Sigma U^{\mathrm{T}}$, $W = \mathrm{diag}(\Sigma)$, 并将每个特征向量 $u_i \in U$ 依据特征值 W 从大到小排列.

Step 4: 计算获得主成分特征序列 $Y = XU$. 其中 $Y = (Y_1, Y_2, \cdots, Y_m)$, Y_1 为第一主成分特征变量, Y_2 为第二主成分特征变量, 依此类推.

通过上述算法, PCA 可以将变量维度为 m 的多元时间序列转变成维度为 K 的主成分特征变量, 即在 Step 3 中提取前 K 个特征向量作为投影坐标系, 原始数据在该坐标系下的投影将形成降维后的特征序列, 即 $Y_{n\times K} = X_{n\times m}U_{m\times K}$. 同时, 通过特征值可以方便地获知降维后多元时间序列数据信息的损失率 $r = 1 - \sum_{k=1}^{K} w_k \Big/ \sum_{i=1}^{m} w_i$; 相反地, 根据给定的损失率 r, 也可获知最大的降维后维度 K.

2 特征表示算法

多元时间序列特征表示的主要目的有两个: 一方面, 通过减少多元时间序列的维度来提高后期数据挖掘的计算效率; 另一方面, 降维后的特征表示能够有效地反映原始时间序列的主要信息. 针对多元时间序列数据, 其高维性表现为时间维度和变量维度, 变量之间的相关性是隐藏在多元时间序列中的重要信息[2,12], 也是多元时间序列相互区分的重要特征之一. 因此, 有必要研究多元时间序列相关性给数据挖掘过程和结果带来的作用和影响.

2.1 相关性特征表示

多元时间序列通常是由特定的系统根据时间的

先后顺序产生的, 不同的系统会产生相异的多元时间序列, 而且系统中各参数的相关性[12]能通过多元时间序列数据上的各个变量来反映. 因此, 若要挖掘系统中各参数之间的关系, 则需要对多元时间序列数据中各变量序列进行相关性分析. 同时, 相关性分析是多元时间序列数据挖掘中重要的特征分析, 可以通过分析每个多元时间序列中变量之间的相关性来反映该时间序列的独特性, 进而判断不同的多元时间序列是否来自于同一系统.

某一多元时间序列数据集 $D = \{D_1, D_2, \cdots, D_L\}$, 对于任意 i 和 j, D_i 和 D_j 表示变量维度为 m 的多元时间序列, 且 D_i 和 D_j 的时间维度可以为任意长度, 即 $n_i \neq n_j$ 或 $n_i = n_j$, 其中 n_i 和 n_j 分别表示 D_i 和 D_j 的时间维度.

由于多元时间序列由系统中的不同参数产生, 各变量序列之间存在量纲, 需要消除各个变量序列的量纲影响, 即 $D = \text{zscore}(D)$. 换言之, 对于任一变量序列 X_i, 其标准化过程为 $X_i = (X_i - \mu_i)/\sigma_i$, 其中 μ_i 和 σ_i 分别为单变量时间序列 X_i 的均值和标准差.

在概率与统计学中, 协方差用于衡量两个变量的总体误差变量, 它可以用来反映两个变量之间的相互关系: 协方差为正值表明两个变量的变化趋势一致; 协方差为负值表明两个变量的变化趋势相反; 协方差为 0 表明这两个变量相互独立. 因此, 通过协方差可以反映系统中各个参数的相关关系, 某一多元时间序列 D_i 的协方差为

$$S_i = \begin{bmatrix} s_{11}^i & s_{12}^i & \cdots & s_{1m}^i \\ s_{21}^i & s_{22}^i & \cdots & s_{2m}^i \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1}^i & s_{m2}^i & \cdots & s_{mm}^i \end{bmatrix}, \tag{2}$$

其中: s_{qp}^i 表示多元时间序列 D_i 中第 q 个变量 X_q^i 与第 p 个变量 X_p^i 之间的协方差, 即 $s_{qp} = \text{E}[(X_q^i - \text{E}(X_q^i))(X_p^i - \text{E}(X_p^i))]$. 由量纲消除过程可得

$$s_{qp} = \text{E}[X_q^i X_p^i]. \tag{3}$$

相关性特征表示 (RFR) 的主要思想是利用每个多元时间序列的协方差矩阵来反映各多元时间序列的信息特征, 通过多元时间序列的变量协方差矩阵描述系统中各参数之间的相互关系, 再通过组合综合协方差矩阵进行奇异值分解, 最终获得基于变量相关性的多元时间序列特征表示. 算法过程如下.

相关性特征表示 $F = \text{RFR}(D)$.

输入: 多元时间序列数据集

$$D = \{D_1, D_2, \cdots, D_L\},$$

其中 $D_i = (X_1^i, X_2^i, \cdots, X_m^i)$;

输出: 特征矩阵 $F = (F_1, F_2, \cdots, F_{mL})$.

Step 1: 对数据集 D 中的每个多元时间序列 D_i 进行标准化处理, 消除各分量之间的量纲影响, 即 $D_i = \text{zscore}(D_i)$.

Step 2: 根据式 (2) 和 (3) 计算数据中每个多元时间序列的协方差矩阵 S_i.

Step 3: 将所有协方差矩阵组合成综合协方差矩阵 Z 来替代原始多元时间序列数据信息, 即 $Z = (S_1, S_2, \cdots, S_L)$. 同时, 对 Z 中的每列去均值处理, 即 $Z_l = Z_l - \bar{Z}_l$, 其中 $\bar{Z}_l$ 表示 Z_l 的均值.

Step 4: 利用 SVD 对综合协方差矩阵 Z 进行主成分计算, 即 $Z^{\text{T}}Z = F\Sigma F^{\text{T}}$.

Step 5: 输出矩阵大小为 $c \times c$ 的特征矩阵 $F_{c\times c}$, 其中 $c = mL$.

与传统 PCA 相比, RFR 是基于全体多元时间序列数据集的特征分析方法, 使得最终获得的特征充分考虑到了其他多元时间序列的信息, 加强了数据序列之间的联系. 同时, RFR 是基于多元时间序列综合协方差的特征分析方法, 比较各个多元时间序列中变量相关性之间的区别和联系, 进而提升各个特征描述原始变量之间相互关系的能力.

特征矩阵 F 表示整个多元时间序列数据集的特征. 在特征集合中, 每 m 个特征向量可视为相应的多元时间序列的特征. 换言之, 在数据集 D 中, 多元时间序列 D_i 经 RFR 特征转化后表示为 $F((i-1)m+1 : im, :)$, 即特征矩阵 F 中从第 $(i-1)m+1$ 行到第 im 行所形成的子矩阵 R. 若要对多元时间序列 D_i 进行降维处理, 则仅需要对 R 保留前 K 列作为 D_i 的特征 R_i, 即 $R_i = F((i-1)m+1 : im, 1 : K)$.

2.2　效率改进与复杂度分析

对数量为 L 的多元时间序列数据集 D 进行主成分分析时, PCA 和 RFR 都事先获得每个多元时间序列的协方差矩阵, 在这方面两者的时间复杂度是相同的. 然而, PCA 需要对每个多元时间序列进行奇异值分解, 得到协方差的特征值和特征向量, 其时间复杂度为 $O(m^3)$. 由此可知, 对于 L 个多元时间序列, PCA 的时间复杂度为 $O(Lm^3)$.

RFR 是基于综合协方差矩阵的主成分分析, 它仅需要对该综合协方差矩阵进行一次特征值和特征向量求解, 其时间复杂度为 $O(c^3)$ (即 $O(L^3m^3)$), $c = Lm$. 与 PCA 相比, $O(L^3m^3) > O(Lm^3)$, 表明原始 RFR (O_RFR) 的时间复杂度大于 PCA. 由 RFR 算法可知, Z 是一个 $m \times c$ 的综合协方差矩阵, 且设 Z 已经过标准化处理, 通过 PCA 对其进行主成分分析, 其

分解过程为 $Z^{\mathrm{T}}Z = U\Sigma U^{\mathrm{T}}$. 由于 $Z^{\mathrm{T}}Z$ 为 $c \times c$ 的矩阵, 其奇异值分解所需要的时间复杂度为 c^3, 且 c 通常非常大, O_RFR 计算效率较低. 然而, 由奇异值分解可得 $ZZ^{\mathrm{T}} = V\Sigma V^{\mathrm{T}}$, 通过矩阵变换和运算法则可得

$$ZZ^{\mathrm{T}}V = V\Sigma \Rightarrow Z^{\mathrm{T}}ZZ^{\mathrm{T}}V = Z^{\mathrm{T}}V\Sigma.$$

令 $B = Z^{\mathrm{T}}V$, 则有

$$Z^{\mathrm{T}}ZZ^{\mathrm{T}}V = Z^{\mathrm{T}}V\Sigma \Rightarrow (Z^{\mathrm{T}}Z)B = B\Sigma.$$

由矩阵特征分解定义和性质可知, B 的施密特正交化矩阵为 $Z^{\mathrm{T}}Z$ 非零特征值对应特征向量组成的矩阵, 即 $\bar{B} = \mathrm{GS}(B)$ 为 $Z^{\mathrm{T}}Z$ 对应非零特征值的特征矩阵, 其中 GS 为 Gram-Schmidt 正交单位化函数. $Z^{\mathrm{T}}Z$ 和 ZZ^{T} 的非零特征值对应的特征矩阵之间的关系为 $\bar{B} = U$, 其中 V 为 ZZ^{T} 的非零特征值对应的特征矩阵, 且有 $B = Z^{\mathrm{T}}V$. 上述推理过程也表明, $Z^{\mathrm{T}}Z$ 和 ZZ^{T} 具有相同的非零特征值. 因此, 大矩阵 $Z^{\mathrm{T}}Z$ 的奇异值分解可以转化成小矩阵 ZZ^{T} 的奇异值分解, 然后通过两者的特征向量之间的关系快速求得大矩阵的特征值和特征向量, 进而实现大矩阵 $Z^{\mathrm{T}}Z$ 的主成分分析.

效率改进算法 $U = \mathrm{FastEig}(Z)$.

输入: 综合协方差矩阵 $Z_{m\times c}$, 其中 $c > m$;

输出: 非零特征值对应的特征矩阵 U.

Step 1: 利用奇异值分解对维度较小矩阵 ZZ^{T} 分解, 即

$$(ZZ^{\mathrm{T}})_{m\times m} = V_{m\times r}\Sigma_{r\times r}V^{\mathrm{T}}_{r\times m}.$$

其中: $\Sigma_{r\times r}$ 为 r 个非零特征值所形成的对角矩阵; V 为 ZZ^{T} 的 r 个非零特征值对应的特征矩阵, 且 V 为 $m \times r$ 的正交矩阵.

Step 2: 计算 $B = Z^{\mathrm{T}}V$, 得到维度较大的矩阵 $Z^{\mathrm{T}}Z$ 非零特征值对应的特征矩阵, 其维度为 $c \times r$, 即对应的 r 个特征列向量.

Step 3: 利用施密特正交化方法对特征矩阵 B 进行正交化处理, 即 $\bar{B} = \mathrm{GS}(B)$.

Step 4: $U = \bar{B}$ 且有 $(Z^{\mathrm{T}}Z)_{c\times c} = U_{c\times r}\Sigma_{r\times r}U^{\mathrm{T}}_{r\times c}$.

由于小矩阵 ZZ^{T} 的维度为 $m \times m$, 对其进行奇异值分解的时间复杂度为 $O(m^3)$, 得到特征矩阵 $V_{m\times m}$. $B = Z^{\mathrm{T}}V$ 的计算时间复杂度为 $O(Lm^2)$, 且 $\mathrm{GS}(B)$ 的时间复杂度为 $O(m^2)$, RFR 对综合协方差矩阵进行奇异值分解的时间复杂度降低到 $O(m^3 + (L+1)m^2)$, 其值远小于原始时间复杂度 $O(L^3m^3)$. 特别地, 当 $m \geqslant \dfrac{L+1}{L^3-1}$ 时, 利用 RFR 进行多元时间序列数据集特征分解的时间复杂度将小于 PCA 所用的时间复杂度. 同时, 数据集中多元时间序列数量 L 越大, RFR 的时间效率相对于 PCA 时间效率的优势越大. 结合计算每个多元时间序列数据协方差矩阵所花费的时间 $O(Lm^2)$, PCA 和 RFR 对多元时间序列数据集 D 的特征表示所需要的时间复杂度如表 1 所示.

表 1 PCA 和 RFR 的时间复杂度

方法	PCA	RFR
时间复杂度	$O(Lm^3 + Lm^2)$	$O(m^3 + (2L+1)m^2)$

3 数值实验

为了更好地理解本文所提出的 RFR 算法并进一步说明该方法的可行性和优越性, 首先, 通过简单实例演示基于 RFR 的金融多元时间序列特征表示; 其次, 比较基于共同主成分分析 CPCA 和 RFR 的拟合协方差; 然后, 通过多元时间序列数据分类实验比较 3 种方法 (PCA、CPCA 和 RFR) 在数据挖掘中的应用效果; 最后, 验证并比较 3 种方法的时间效率.

3.1 实 例

利用 3 支上证指数 (X_1 上证金融地产行业指数 (000038)、X_2 上证金融地产行业分层等权重指数 (000076) 和 X_3 上证信息技术行业分层等权重指数 (000077)) 作为实例演算数据, 各个指数分别由 4 个变量属性构成, 即开盘价、最高价、最低价和收盘价, 其截取时间为 2012-05-01 ∼ 2013-04-30, 这 3 支上证指数形成了维度为 242×4 的多元时间序列 $X = \{X_1, X_2, X_3\}$. 由 RFR 算法 Step 1 ∼ Step 3 可以得到综合协方差矩阵

$$Z = \left[\begin{array}{cccc} 1.000\,0 & 0.994\,2 & 0.996\,8 & 0.989\,8 \\ 0.994\,2 & 1.000\,0 & 0.995\,5 & 0.997\,5 \\ 0.996\,8 & 0.995\,5 & 1.000\,0 & 0.995\,1 \\ 0.989\,8 & 0.997\,5 & 0.995\,1 & 1.000\,0 \end{array}\right] \rightarrow$$

$$\leftarrow \left[\begin{array}{cccc} 1.000\,0 & 0.993\,0 & 0.996\,2 & 0.987\,8 \\ 0.993\,0 & 1.000\,0 & 0.994\,5 & 0.997\,0 \\ 0.996\,2 & 0.994\,5 & 1.000\,0 & 0.994\,0 \\ 0.987\,8 & 0.997\,0 & 0.994\,0 & 1.000\,0 \end{array}\right] \rightarrow$$

$$\leftarrow \left[\begin{array}{cccc} 1.000\,0 & 0.995\,9 & 0.995\,0 & 0.986\,8 \\ 0.995\,9 & 1.000\,0 & 0.996\,2 & 0.995\,0 \\ 0.995\,0 & 0.996\,2 & 1.000\,0 & 0.995\,4 \\ 0.986\,8 & 0.995\,0 & 0.995\,4 & 1.000\,0 \end{array}\right].$$

其中: 每 4 列所形成的矩阵就是相应多元时间序列的协方差矩阵. 同时, Z 通过均值化处理后根据 Step 4 利用 SVD 对其进行特征分解, 便可得到维度为 12×12 的特征矩阵 F, 每一列对应一个特征向量, 且这些特征向量按照特征值的大小排列. 若需要降维处理且特征维度 $k = 2$, 则从特征矩阵 F 中选取前两列作为

综合协方差矩阵 Z 的特征 $R = U(:, 1:2)$, 从而有

$$R^{\mathrm{T}} = \left[\begin{array}{cccc} -0.344 & 0.139 & -0.083 & 0.347 \\ -0.033 & 0.336 & -0.345 & 0.032 \end{array}\right| \rightarrow$$
$$\leftarrow \left.\begin{array}{cccc} -0.413 & 0.170 & -0.103 & 0.415 \\ -0.050 & 0.414 & -0.425 & 0.041 \end{array}\right| \rightarrow$$
$$\leftarrow \left.\begin{array}{cccc} -0.415 & -0.001 & -0.010 & 0.425 \\ 0.332 & 0.357 & -0.317 & -0.269 \end{array}\right] =$$
$$(R_1^{\mathrm{T}} | R_2^{\mathrm{T}} | R_3^{\mathrm{T}}).$$

其中: R_1, R_2 和 R_3 分别表示原始 3 支金融多元时间序列 X_1, X_2 和 X_3 的相关性特征, 从各自变量相关性角度出发, 研究所有多元时间序列之间在变量相关性方面的共性. 就数据特征而言, R_1 与 R_2 数据分布比较近似, 例如, R_1 与 R_2 的正负值分布在相同位置上, 并且数据之间比较接近. 另外, 通过欧氏距离来度量这些特征之间的相似性, 其距离矩阵为

$$d = \begin{bmatrix} 0 & 0.003\,0 & 0.032\,6 \\ 0.003\,0 & 0 & 0.036\,8 \\ 0.032\,6 & 0.036\,8 & 0 \end{bmatrix}.$$

由距离矩阵 d 可知, X_1 与 X_2 更相似. 若结合层次聚类算法, 则 X_1 事先和 X_2 聚类在一个簇中. 同时, 也说明了基于 RFR 的金融多元时间序列聚类分析结果符合客观事实, X_1 和 X_2 的成分股是银行和地产类等行业指数, 而 X_3 为电子、信息和软件等行业指数. 从实例中易发现, 原来维度为 242×4 的多元时间序列 X_i 可用维度为 2×4 的特征矩阵 R_i^{T} 表示, 这不仅有效地实现了数据降维, 而且可使其特征矩阵较好地区分原始数据之间的相异性.

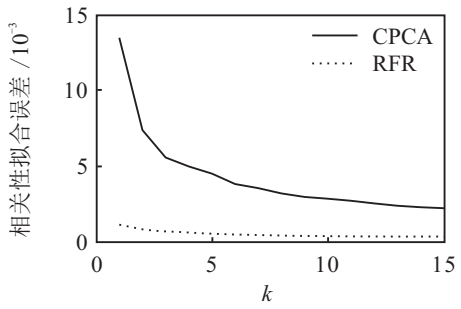
3.2　相关性拟合误差

由 2.1 节分析可知, 协方差矩阵能够很好地表达多元时间序列各分量序列之间的相关关系. 在传统方法中, 基于共同主成分分析的多元时间序列特征表示(CPCA)[10]利用所有数据对象的均值协方差矩阵进行主成分分析, 进而形成统一的投影空间, 并将每个多元时间序列在此空间的投影所形成的数据作为相应的主成分. 同样, RFR 方法也是基于协方差矩阵, 并且它综合考虑了所有协方差矩阵的共性特征. 本次实验比较了两种基于协方差矩阵的相关性拟合误差.

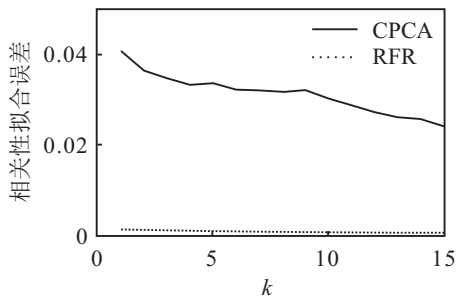
实验数据为 EEG 和 ASL 两种多元时间序列数据集. ASL 具有 22 个变量, 分别表示手和脚的动作特征以及各手指的弯曲程度, 该数据集包含 95 种语意(95 个类), 每种语意有 27 组序列. 选取前 8 种语意对应的序列作为实验数据集, 一共 216 个样本, 样本长度不等且在 47～95 之间. EEG 具有 256 个时间点和 64 个部位(变量)所形成的脑电图数据, 不失一般性, 选取编号为 co2a0000364 和 co2c0000337 两种病人的 10 次测试作为实验数据集, 一共 20 个多元时间序列.

利用 CPCA 和 RFR 分别对多元时间序列数据集 EEG 和 ASL 进行特征表示, 并根据降维力度建立相应的数据投影空间, 利用多元时间序列重构方法 $Y = XU(:, 1:k)U(:, 1:k)^{\mathrm{T}}$ 可以得到相应的拟合序列, 其中 $U(:, 1:k)$ 表示前 k 个特征向量所组成的特征矩阵. 根据式 (2) 计算基于拟合序列的协方差矩阵 S^Y 和基于原始序列的协方差矩阵 S^X, 通过两个协方差矩阵之间的误差来衡量这两种方法在描述多元时间序列变量之间相关性的能力.

CPCA 和 RFR 两种方法在两个数据集 EEG 和 ASL 上的相关性拟合误差如图 1 所示.



(a) 在 EEG 数据集的相关性拟合误差



(b) 在 ASL 数据集的相关性拟合误差

图 1　两种方法在两个数据集中的相关性拟合误差

在图 1 中: 根据不同降维后的维度, 利用 CPCA 和 RFR 对两种多元时间序列数据集进行特征表示并重构数据, RFR 得到的协方差误差小于 CPCA 产生的协方差误差, 表明 RFR 在进行多元时间序列特征表示时, 更能够充分考虑到变量之间的相关性问题, 并以变量相关性作为多元时间序列的主要特征来区分数据之间的相异性.

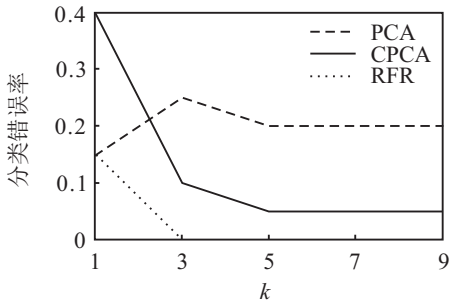
3.3　分类实验

分类是数据挖掘中最经典的方法之一, 它能很好地检验特征表示的有效性. 本次实验利用最近邻分类算法对前一节所提到的数据集 EEG 和 ASL 进行分类, 即首先利用 3 种方法 PCA、CPCA 和 RFR 对数据集进行特征表示, 再将各自的特征集同时视为训练集

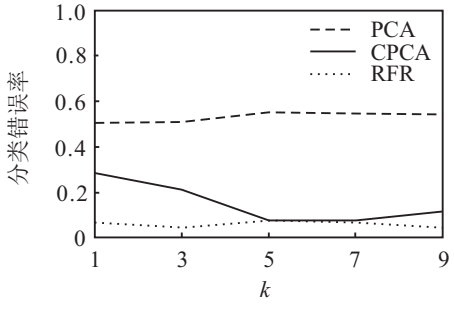
和测试集, 并利用最近邻分类算法从训练集中找出测试集中每个特征序列的最相似对象 (最相似的对象为查询对象本身). 若查询结果的类标与查询对象的类标一致, 则视为正确分类; 否则, 视为错误分类. 最终计算查询结果的平均值, 并比较它们在不同降维维度下的分类结果.

另外, 由于PCA和CPCA所得到的主成分特征序列的长度受原始多元时间序列长度的影响, 若原始多元时间序列数据集中各个序列的时间长度不相等, 则得到的主成分特征序列的长度也不相等, 故在进行分类实验中需要借助不等长序列的距离度量函数动态时间弯曲 (DTW) 来计算特征序列之间的相似度. 在EEG数据集中, 各多元时间序列长度相等, 使得各特征序列长度也彼此相等, 故借助欧氏距离便可计算其相似性; 在ASL数据集中, 各多元时间序列长度不相等, 在分类过程中采用DTW计算相似性. 就RFR而言, 通过数据降维, 可以将不等长的多元时间序列转化为长度相等的相关性特征向量, 故仅用欧氏距离便可计算它们之间的相似性.

3种方法根据不同降维维度在两个数据集中进行分类的结果如图2所示. 在图2(a)中, PCA和CPCA的分类结果是建立在DTW相似性度量基础上的分类结果, 图2(b)的分类结果是建立在欧氏距离上的分类结果. 由实验结果易知: 1) 不管在哪个降维后维度, 基于RFR的特征表示都能得到最好的分类结果; 2) RFR所得到的特征序列具有相同的长度, 解决了由传统方法PCA和CPCA产生不等长特征序列而导致的问题.



(a) EEG数据集的分类结果



(b) ASL数据集的分类结果

图 2 3种方法在两个数据集中的分类结果比较

3.4 时间效率分析

时间效率是判断多元时间序列数据挖掘算法性能的另外一项重要指标. 利用3种方法 (PCA、CPCA和RFR) 对ASL数据集中的多元时间序列进行特征表示, 记录其计算时间代价. 本次实验主要检验3种方法与数据集中多元时间序列数量之间的关系, 故在ASL数据集中按照L的大小进行随机分组, 每组被视为一个数据子集, 考查在数量大小不同的数据子集下3种方法的计算时间代价, 其中L的值为 (2, 22, 42, 62, 82, 102, 122, 142, 162, 182, 202). 通过利用Matlab 7.10在Intel(R) Core(TM)i5-2520M 2.5 GHz且内存为4 GB的64位win 7操作系统环境下运行程序, 3种方法对不同数据子集中多元时间序列特征表示的计算时间代价结果如图3所示.

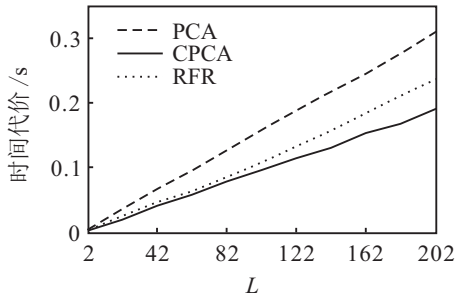


图 3 3种方法在不同数据子集下的特征表示计算时间代价

由图3结果比较可知, PCA需要最多的时间代价, 其次是RFR, 而CPCA的运行时间最少. 实事上, PCA由于需要对数据子集中的每个多元时间序列进行主成分分析, 需要多次计算特征值和特征向量, 会消耗最多的计算时间. 与PCA相比, CPCA和RFR两种方法都是同时考虑数据集中所有多元时间序列的方法, 且这两种方法仅需要计算一次特征值和特征向量, 因此它们的时间代价要小于PCA的运行时间. 另外, 由于CPCA是对大小为$m \times m$的平均协方差矩阵计算特征值和特征向量, 而RFR是对$m \times Lm$的综合协方差矩阵进行特征值和特征向量计算, RFR的时间代价会高于CPCA但小于PCA.

需要说明的是, 本次实验是基于第2.2节中提出的效率改进方法. 若相关性特征表示方法中直接对综合协方差Z进行特征值和特征向量计算, 则O_RFR会因Z的列向量个数Lm过大而增大特征值和特征向量的求解时间, 即L的增大使得求解时间呈三阶指数增长. 然而, 通过效率改进方法大大提高了相关性特征表示方法的计算性能.

相关性特征表示方法在效率改进前后的计算时间代价如图4所示.

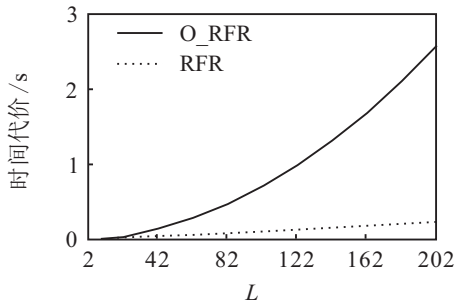


图 4　相关性特征表示方法在效率改进前后的计算时间代价

在图 4 中, 原始相关性特征表示方法 O_RFR 随着 L 的增大呈指数性增长, 相比而言, 改进后的 RFR 具有较好的计算性能.

4　结　论

基于多元时间序列各分量属性之间关系的重要性, 本文从多元时间序列变量相关性的角度出发, 通过构建综合协方差矩阵来共同反映原始数据集中各多元时间序列之间的关系; 然后, 利用奇异值分解方法对综合协方差矩阵进行主成分分析, 进而得到反映原始多元时间序列变量相关性的特征序列. 与传统方法相比, 该方法具有以下几点优势: 1) 利用变量相关性描述原始多元时间序列的数据特征, 使最终的特征表示能够反映原始多元时间序列中变量之间相关性的独立性以及与其他多元时间序列的联系性, 提高了特征表示结果对数据关系的描述能力; 2) RFR 方法所得到的特征序列具有相同的长度, 解决了由传统方法 PCA 和 CPCA 产生不等长特征序列导致的问题; 3) 提出了相应的效率改进算法, 仅需要计算维度较小的矩阵的非零特征值对应的特征矩阵, 便可快速有效地得到维度较大矩阵的非零特征值对应的特征矩阵; 4) RFR 能够用最少的特征来表示原始多元时间序列的相关信息, 并且其特征表示能够提高后期多元时间序列数据挖掘的质量.

参考文献(References)

[1] Esling P, Agon C. Time-series data mining[J]. ACM Computing Surverys, 2012, 45(1): 11-12.

[2] Wang Zhenxing, Chan Laiwan. Learning causal relations in multivariate time series data[J]. ACM Trans on Int Systems and Technology, 2012, 3(4): 71-76.

[3] Li Hailin, Yang Libin, Guo Chonghui. Improved piecewise vector quantized approximation based on normalized time subsequences[J]. Measurement, 2013, 46(9): 3429-3439.

[4] 闫秋艳, 夏士雄. 一种无限长时间序列的分段线性拟合算法[J]. 电子学报, 2010, 38(2): 443-448.
(Yan Q Y, Xia S X. An piecewise linear fitting algorithm for infinite time series[J]. Acta Electronica Sinica, 2010, 28(2): 443-448.)

[5] 李海林, 杨丽彬. 时间序列数据降维及特征表示新方法[J]. 控制与决策, 2013, 28(11): 1718-1722.
(Li H L, Yang L B. Method of dimensionality reduction and feature representation for time series[J]. Control and Decision, 2013, 28(11): 1718-1722.)

[6] Yang K, Shahabi C. An efficient k nearest neighbor search for multivariate time series[J]. Information and Computation, 2007, 205(1): 65-98.

[7] Weng X, Shen J. Classification of multivariate time series using two-dimensional singular value decomposition[J]. Knowledge-Based Systems, 2008, 21(7): 535-539.

[8] 韩敏, 李德才. 基于 EOF-SVD 模型的多元时间序列相关性研究及预测[J]. 系统仿真学报, 2008(7): 1669-1672.
(Han M, Li D C. Multiple time series correlation extraction and prediction based on EOF-SVD model[J]. J of System Simulation, 2008(7): 1669-1672.)

[9] Baragona R, Battaglia F. Outliers detection in multivariate time series by independent component analysis[J]. Neural Computation Archive, 2007, 19(7): 1962-1984.

[10] 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.)

[11] Keogh E. Exact indexing of dynamic time warping[J]. Knowledge and Information Systems, 2005, 7(3): 358-386.

[12] Li Hailin. Asynchronism-based principal component analysis for time series data mining[J]. Expert Systems with Applications, 2014, 41(6): 2842-2850.

[13] 王文波, 张晓东, 汪祥莉. 基于主成分分析的经验模态分解消噪方法[J]. 电子学报, 2013, 41(7): 1425-1430.
(Wang W B, Zhang X D, Wang X L. Empirical mode decomposition de-noising method based on principal component analysis[J]. Acta Electronic Sinica, 2013, 41(7): 1425-1430.)

[14] 姜慧研, 宗茂, 刘相莹. 基于 ACO-SVM 的软件缺陷预测模型的研究[J]. 计算机学报, 2011, 34(6): 1148-1154.
(Jiang H Y, Zong M, Liu X Y. Research of software defect prediction model based on ACO-SVM[J]. Chinese J of Computers, 2011, 34(6): 1148-1154.)

[15] 林琳, 陈虹, 陈建. 基于鲁棒听觉特征的说话人识别[J]. 电子学报, 2013, 41(3): 619-624.
(Lin L, Chen H, Chen J. Speaker recognition based on robust auditory feature[J]. Acta Electronic Sinica, 2013, 41(3): 619-624.)

(责任编辑: 闫　妍)