计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(11):153−160 [doi: 10.15888/j.cnki.csa.007139]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 2DSVD 的多变量时间序列半监督分类①

单中南, 翁小清, 马超红

(河北经贸大学 信息技术学院, 石家庄 050061)
通讯作者: 单中南, E-mail: 1027751712@qq.com

摘　要: 目前时间序列半监督分类研究主要集中在单变量时间序列, 由于多变量时间序列 (MTS) 变量之间存在复杂关系, MTS 的半监督分类研究比较少. 针对这种情况, 提出一种基于二维奇异值分解的 MTS 半监督分类方法, 该方法首先计算行-行以及列-列协方差矩阵的特征向量, 然后从 MTS 样本中提取特征矩阵; 特征矩阵的行数以及列数不仅比原 MTS 样本低, 而且还清晰地考虑了 MTS 样本的二维特性. 在 10 个 MTS 数据集上的实验结果表明, 该方法的分类性能显著地好于使用扩展 Frobenius 范数、中心序列、以及基于一维奇异值分解的半监督分类方法.

关键词: 多变量时间序列; 二维奇异值分解 (2DSVD); 半监督分类

引用格式: 单中南,翁小清,马超红.基于 2DSVD 的多变量时间序列半监督分类.计算机系统应用,2019,28(11):153–160. http://www.c-s-a.org.cn/1003-3254/7139.html

Semi-Supervised Classification of Multivariate Time Series Based on Two-Dimensional Singular Value Decomposition

SHAN Zhong-Nan, WENG Xiao-Qing, MA Chao-Hong

(Information Technology College, Hebei University of Economics and Business, Shijiazhuang 050061, China)

Abstract: At present, semi-supervised classification research of time series mainly focuses on univariate time series, due to the complex relationship between Multivariate Time Series (MTS) variables, there is less research on semi-supervised classification of MTS. In view of this, we proposes a semi-supervised MTS classification method based on Two-Dimensional Singular Value Decomposition (2DSVD), which first computes the eigenvectors of row-row and column-column covariance matrices, and then extracts feature matrices from MTS samples. The number of rows and columns of the feature matrix is not only lower than the original MTS sample, but also clearly considers the two-dimensional nature of the MTS sample. The experimental results on 10 MTS datasets show that the semi-supervised classification performance of this method is significantly better than the method using extended Frobenius norm, center sequence, and based on one dimensional singular value decomposition.

Key words: Multivariate Time Series (MTS); two-Dimensional Singular Value Decomposition(2DSVD); semi-supervised classification

时间序列是指按时间次序有序排列的一组数据, 任何有次序的实值序列都可当作时间序列来处理[1]. 时间序列数据广泛地存在于金融、医学、交通等领域. 建立准确的分类器需要大量的有类别标记的样本数据, 然而在现实应用领域, 存在大量没有类别标记的样本数据, 有标记的样本数据很难获得, 或用人工标记样本数据成本很高. 半监督分类 (Semi-Supervised Classification, SSC) 使用少量有标记的样本数据和大量未标记

① 收稿时间: 2019-04-03; 修改时间: 2019-05-08; 采用时间: 2019-05-13; csa 在线出版时间: 2019-11-06

的样本数据建立分类器. 目前, 绝大多数时间序列半监督分类的研究工作都集中在单变量时间序列, 对多变量时间序列 (Multivariate Time Series, MTS)的半监督分类研究还比较少. 在对 MTS 进行半监督分类时, 主要遇到两方面的困难[2]: 第一, MTS 中含有多个变量, 且变量之间存在复杂的相关系; 第二, 不同 MTS 样本它们的长度不一定相等, 这些困难使得标准的分类器很难直接使用. 本文针对 MTS 特性, 采用二维奇异值分解 (Two-Dimensional Singular Value Decomposition, 2DSVD) 从 MTS 样本中提取特征矩阵, 并与其他 MTS 半监督分类方法进行性能对比, 讨论该方法在 MTS 半监督分类的优势. 本文第 1 节介绍背景和相关工作; 第 2 节提出了基于 2DSVD 的 MTS半监督分类算法; 第 3 节通过实验将本文提出的方法与其它半监督分类方法进行比较, 并采用威尔克森符号秩检验 (Wilcoxon signed ranks test) 对实验结果进行对比, 验证算法的有效性; 第 4 节给出了本文结论.

1 背景和相关工作

1.1 基本概念

定义 1. 时间序列. 时间序列是一段时间内的一系列观测值, 用 $x_i(t)[i=1, 2, \cdots, n; t=1, 2, \cdots, m]$表示, 其中 m 是观测值的个数, n 是变量的个数[2]. 当 $n=1$ 时, 称为单变量时间序列, 当 $n\geq 2$ 时, 称为多变量时间序列, 通常用 $m\times n$ 矩阵存储 MTS.

定义 2. P 集合. P 为训练数据的一个集合, 包括所有正类标记的样本[3]. 在训练开始时, P 只包含少量的正类样本, 或许只包含一个正类样本. 随着学习的继续, 先前 U 中一些没有标记的样本, 被标记为正类样本, 并移动到了 P 集合, P 集合包含样本的数量也随之增加. 最终, 集合 P 既包含原来有标记的正类样本, 也包括使用分类器从 U 中选择的样本.

定义 3. U 集合. U 是未标记样本的集合[3]. U 中的样本可以来自正类或者负类; 通常情况下, U 中的绝大多数样本来自负类.

1.2 二维奇异值分解

Ding 等[4]对标准奇异值分解 (即一维奇异值分解, 1DSVD) 进行了扩展, 提出了基于行-行协方差矩阵以及列-列协方差矩阵的二维奇异值分解方法, 2DSVD 是基于二维矩阵而不是基于一维向量[2]. 2DSVD 使用 MTS 样本构造行-行以及列-列的协方差矩阵, 然后计算行-行及列-列协方差矩阵的特征向量用于 MTS 样本特征矩阵的提取. 使用 2DSVD 提取出的 MTS 样本的特征矩阵, 它们的行数以及列数不仅比原始数据低, 而且还清晰地考虑了原始数据的二维特性.

设 $\{T_i\}_{i=1}^N$ 是一个由 MTS 样本构成的集合, 其中 $T_i\in R^{m\times n}$, m 为观测值的个数, n 为变量的个数, N 为集合中样本的个数. MTS 行-行协方差矩阵 F 以及列-列协方差矩阵 G 定义如下[2,5]:

$$F=\frac{1}{N}\sum_{i=1}^{N}(T_i-\overline{T})(T_i-\overline{T})^{\mathrm{T}} \tag{1}$$

$$G=\frac{1}{N}\sum_{i=1}^{N}(T_i-\overline{T})^{\mathrm{T}}(T_i-\overline{T}) \tag{2}$$

其中, $\overline{T}=\sum_i T_i/N$. 设 U_r 包含行-行协方差矩阵 F 的前 r 个主要特征向量, $U_r=(u_1,\cdots,u_r)$; V_s 包括列-列的协方差矩阵 G 的前 s 个主要特征向量, $V_s=(v_1,\cdots,v_s)$. MTS 样本集合 $\{T_i\}_{i=1}^N$ 的 2DSVD 表示为: $\left\{U_r,V_s,\{M_i\}_{i=1}^N\right\}$, 其中 $M_i=U_r^{\mathrm{T}}T_iV_s$, $U_r\in R^{m\times r}$, $V_s\in R^{n\times s}$, $M_i\in R^{r\times s}$, M_i 即为从 MTS 样本 T_i 中提取出的特征矩阵. 设 $M_i=\left[Y_1^i,Y_2^i,\cdots,Y_s^i\right]$为从 MTS 样本 $T_i(i=1, 2)$ 中提取的特征矩阵, 特征矩阵 M_1 与 M_2 之间的距离定义为:

$$d(M_1,M_2)=\sum_{k=1}^{s}\left\|Y_k^1-Y_k^2\right\| \tag{3}$$

其中, $\|\cdot\|$为 L2 范数.

1.3 相关工作

时间序列的半监督分类方法可大致分为 3 类[6,7]: 基于实例、基于聚类以及基于模型的半监督分类方法.

Wei 等[8]针对正类中只有少量有标记的样本, 使用欧氏距离建立基于最小最近邻距离的分类器及停止准则. Ratanamahatana[9]等使用 DTW (Dynamic Time Warping) 距离来改进样本的选取并提出了新的停止准则, 该准则基于未标记样本集中候选样本与正类样本的历史距离; Chen[3]等在 SSC 算法中, 使用一种基于 DTW 和 ED 相结合的特殊距离 DTW-D, 显著地提高了分类的性能. Begum 等[10,11]提出了一种基于最小描述长度 (Minimum Description Length, MDL) 的停止准则, 该准则利用数据的内在性质去发现停止点; 然而, 时间序列在时间轴可能会存在扭曲 (distortion) 现象, 出现不匹配点, Vinh 等[12,13]针对此问题进行了改进, 并

增加一个后处理步骤, 使分类器更加精确. Vinh 等[14]还提出了一种基于约束的自训练算法, 与正类集合最近的实例 t, 必须满足约束 DL(t|H)<DL(t), 才能添加到正类集合. 另外, Vinh 等还定义了安全距离 (safe distance), 当实例与正类集合之间的距离小于或等于安全距离, 则将该实例放入正类集合中.

目前绝大多数研究工作集中在单变量时间序列半监督分类算法性能的提高, 以及停止准则的改进方面, 对 MTS 半监督分类的研究很少. 在对 MTS 进行半监督分类时, 主要存在变量之间的复杂相关关系以及样本长度不一致等因素, 使得标准分类器很难直接使用. Li 等[15,16]提出了两种基于标准 SVD 的特征提取方法 (以下简称 Li's first、Li's second 方法) 用于 MTS 分类, Li's first 方法是将第 1 个奇异向量 u_1 与经过标准化后由奇异值组成的向量 $\sigma_{\text{normalized}}$ 相连, 作为 MTS 样本的特征表示. Li's second 方法将加权以后的第 1 奇异向量 w_1u_1 与加权后的第 2 奇异向量 w_2u_2 相连, 作为 MTS 样本的特征表示. 这两种方法本质上属于一维奇异值分解, 但是 MTS 包含变量维和时间维两个维度, 本文提出基于 2DSVD 的半监督分类方法, 从行和列两个方向对 MTS 样本进行降维, 清晰地考虑了 MTS 样本的二维特性.

2 基于 2DSVD 的 MTS 半监督分类算法

2.1 训练分类器

本文提出的 MTS 半监督分类算法主要包括 4 个步骤: 第一步, 使用未标记数据集 U 来计算变换矩阵 U_r 以及 V_s, 获取每个训练样本的特征矩阵; 第二步, 随机选取若干个正类样本的特征矩阵作为初始标记数据 P; 第三步, 计算集合 U 中每个样本到集合 P 的欧氏距离, 将集合 U 中与集合 P 最近的样本, 从集合 U 中删除, 添加至集合 P; 第四步, 重复第三步, 直到满足停止标准为止.

基于 2DSVD 的 MTS 半监督分类算法如算法 1 所示. 在步骤 7 中, 本文采用 Wei 等[8]提出的停止标准, 即在迭代过程中, 当正类样本的最小最近邻距离在趋于稳定后的第一次显著下降时, 即停止. TWOSVDSSC 分为两个阶段, 步骤 1–步骤 5 为降维阶段: 设未标记数据集 U 中有 M 个 MTS 样本, 算法的行-行协方差矩阵 F 为 $m\times m$ 矩阵, 列-列协方差矩阵 G 为 $n\times n$ 矩阵[5], 由于对 $n\times n$ 矩阵进行奇异值分解的时间复杂度为 O(n^3)[2], 所以算法中步骤 1–步骤 4 的时间复杂度为 O(m^3+n^3); 步骤 5 是计算未标记数据集 U 中每一个 MTS 样本的特征矩阵, 时间复杂度为 O(M*r*s), 由于在 MTS 样本中, 变量个数 n 以及参数 r 和 s 往往都远小于样本长度 m, 因此步骤 1–步骤 5 的时间复杂度主要取决于样本长度; 步骤 6–步骤 8 为训练分类器阶段, 时间复杂度为 O(M^2). 所以算法的复杂度为 O((m^3+n^3)+(M*r*s)+M^2).

分类器训练好之后, 在使用分类器对待测样本进行分类时, 如果待测样本与任何一个标记为正类样本之间的距离小于阈值 r, 则该样本分类为正类, 否则为负类[8], 阈值 r 为正类样本与其最近邻之间距离的平均值.

算法 1. 基于 2DSVD 的 MTS 半监督分类算法

输入: P 是初始训练集, 包含少量已标记正类样本; U 是未标记数据集; *nSeeds* 是初始标记为正类样本的个数.
输出: 训练好的分类器.

1. 计算 U 中行-行协方差矩阵 F;
2. 使用 SVD 计算 F 的特征向量, 由 F 的前 r 个主要特征向量组成的变换矩阵 U_r;
3. 计算 U 中列-列协方差矩阵 G;
4. 使用 SVD 计算 G 的特征向量, 由 G 的前 s 个主要特征向量组成的变换矩阵 V_s;
5. 计算 U 中每个 MTS 样本的特征矩阵 M_i;
6. 随机选取 *nSeeds* 个正类样本放入集合 P;
7. 计算集合 U 中每个样本到集合 P 的欧氏距离, 将集合 U 中与集合 P 最近的样本, 从集合 U 中删除, 添加至集合 P;
8. 重复步骤 7, 直到满足停止标准为止.

2.2 评估分类器

算法 1 仅包含来自 U 中的正类样本, 属于一类分类器. 本文采用测试集对分类器的性能进行测试, 测试集中包含一些正类样本和其他类样本. 采用经典的精确度 (*Precision*) 和召回率 (*Recall*) 来衡量分类器的性能. 在本文中, 精确度的值等于召回率的值, 即假的负类 (False negatives) 数量与假的正类 (False positives) 数量相同. 精确度的定义如下所示[3], 其中 K 是指测试集中的正类样本的个数, N_{positive} 为在前 K 个最接近 P 集合的样本中, 正类样本的个数.

$$Precision = \frac{N_{\text{positive}}}{K} \tag{4}$$

3 实验

3.1 数据集描述

本文实验数采用的 Lp1、Lp2、Lp4、Lp5 数据

集[17]包含机器人在故障检测后的力和扭矩测量值. 每个故障的特征是在故障检测后每隔一段时间收集的 15 个力/扭矩样本, Lp1、Lp2、Lp4、Lp5 数据集中每个样本包含 6 个变量; BCI 数据集[18,19]中 MTS 样本分为两种类型: 一种是被测试者用左手手指按计算机键盘时的脑电图 (EEG) 情况, 有 208 个样本; 另一种是被测试者用右手手指按计算机键盘时的脑电图情况, 也有 208 个样本. 数据集中每个样本包含 28 个变量; Japanese Vowels 数据集[20]记录 9 个男性在发日语的元音/ae/, 这 9 个男性对应的样本个数分别为: 61, 65, 118, 74, 59, 54, 70, 80 以及 59, 数据集中每个样本包含 12 个变量; Wafer 数据集[21]记录真空室传感器监控半导体微电子的制造过程, 每一个硅晶片的生产过程可以用含有 6 个变量的 MTS 样本来描述, 并被分为正常或异常两类, 数据集中包含 327 个 MTS 样本并被分为 2 类: 其中正常样本有 200 个, 异常样本有 127; AUstralian Sign LANguage(以下简称 AUSLAN) 数据集[20]由随机选取 25 种手势的 MTS 样本 (总共 675 个 MTS 样本) 组成, 每个样本包含 22 个变量; Character Trajectories 数据集[22]中所有样本来自同一位作者, 通过书写单个字符来记录笔尖 (pen tip) 轨迹, 记录时只考虑带有单一落笔段的字符, 每个样本包含 x 和 y 坐标以及笔尖力度这 3 个变量; Gas sensors 数据集[23,24]包含由 MOX 以及温度和湿度这三种传感器组成的气体传感器, 记录来自 3 种不同气体所产生的观测值, 数据集中每个样本包含 10 个变量. 表 1 列出了 10 个 MTS 数据集的主要特征. 2DSVD 要求数据集中所有 MTS 样本具有相同长度. 对于具有不同长度样本的 MTS 数据集, 本文采用 Rodriguez 等[25]提出的方法, 将所有 MTS 样本的长度都延长到该数据集中最长 MTS 样本的长度. 延长方法如下: 如将长度为 100 的 MTS 样本延长至 120, 只需将样本中每 5 个值中的一个值复制即可. 该方法使得原样本中的所有值都保留在延长后的样本中, 不会损失任何数据信息.

表 1 数据集描述

数据集名称	变量个数	最大长度	最小长度	类别个数	样本总数
Lp1	6	15	15	4	88
Lp2	6	15	15	5	47
Lp4	6	15	15	3	117
Lp5	6	15	15	5	164
BCI	28	500	500	2	416
Japanese Vowels	12	29	7	9	640
Wafer	6	198	104	2	327
AUSLAN	22	95	47	25	675
Character	3	205	109	20	2858
Gas sensors	10	15 393	3825	3	99

3.2 性能比较

将本文提出的基于 2DSVD 的 MTS 特征提取方法, 与基于扩展 Frobenius 范数的距离 D_{Eros}[26]、中心序列[27]、以及基于一维 SVD 的 Li's first, Li's second 方法[15,16]分类性能进行比较. 在实验中, 将数据集中类别标记为 1(class label=1) 的样本选为正类样本数据, 其它类样本皆为负类样本数据. 在算法 2.1 中, 初始正类样本的个数 *nSeeds* 分别取 1、3、5 个, 实验重复 100 次, 表 2、3、4 给出了各种方法 100 次实验的平均 *Precision*.

表 2、表 3、表 4 给出了在 10 个数据集上使用不同方法进行半监督分类的 *Precision*. 表中列 2 和列 3 给出了在数据集上使用基于扩展 Frobenius 范数的距离 D_{Eros}[26]以及中心序列[27]的方法进行分类的 *Precision*; 表中列 4 和列 5 给出了在数据集上使用 Li's first 以及 Li's fecond 方法进行分类的 *Precision*; 列 6 给出了使用 2DSVD 进行分类时最高的 *Precision* 以及相应参数 r 和 s 的值, 其中, r 和 s 分别表示使用 2DSVD 方法得到对应特征矩阵的行及列的个数.

从表 2 可以看出, 当初始正类样本的个数 *nSeeds* 为 1 时, 2DSVD 在 10 个 MTS 数据集上分类的平均 Precision 为 0.76, D_{Eros} 的平均值为 0.39, 中心序列的平均值为 0.63, Li's First 以及 Li's Second 的平均值分别为 0.53 和 0.52; 从表 5 中可以看到, 2DSVD 与

其它 4 种方法的 Wilcoxon 符号秩检验的概率 p 值都小于 0.05, 说明 2DSVD 的分类性能显著地好于其它四种方法. 当 *nSeeds* 为 3 或 5 时, 也可以得到相同的结论. 从表 2、表 3、表 4 中还可以看出, 各种方法的平均 *Precision* 随着 *nSeeds* 增大而增大, 说明增加初始正类样本个数, 能够提高算法的分类性能.

表 2 *nSeeds*=1 时各种方法的 *Precision*

数据集	D_{Eros}	中心序列	Li's first	Li's second	2DSVD
Lp1	0.79	1.00	1.00	1.00	1.00(r=7, s=3)
Lp2	0.54	0.67	0.77	0.72	0.97(r=15, s=1)
Lp4	0.47	0.76	0.34	0.34	0.96(r=15, s=6)
Lp5	0.36	0.95	0.53	0.54	0.90(r=15, s=5)
BCI	0.52	0.46	0.47	0.47	0.47(r=500, s=2)
Vowel	0.18	0.55	0.73	0.73	0.70(r=1, s=12)
Wafer	0.30	0.39	0.25	0.26	0.47(r=28, s=1)
AUSLAN	0.45	0.28	0.74	0.75	0.88(r=1, s=21)
Character	0.15	0.78	0.26	0.22	0.80(r=5, s=3)
Gas sensors	0.17	0.44	0.20	0.21	0.45(r=170, s=2)
平均值	0.39	0.63	0.53	0.52	0.76

表 3 *nSeeds*=3 时各种方法的 *Precision*

数据集	D_{Eros}	中心序列	Li's first	Li's fecond	2DSVD
Lp1	0.82	1.00	1.00	1.00	1.00(r=7, s=3)
Lp2	0.71	0.82	0.95	0.92	0.99(r=15, s=1)
Lp4	0.59	0.84	0.58	0.54	0.99(r=15, s=6)
Lp5	0.46	0.95	0.54	0.54	0.97(r=15, s=5)
BCI	0.51	0.46	0.45	0.46	0.47(r=500, s=2)
Vowel	0.19	0.53	0.79	0.77	0.78(r=1, s=12)
Wafer	0.22	0.40	0.11	0.15	0.49(r=28, s=1)
AUSLAN	0.40	0.31	0.81	0.84	0.91(r=1, s=21)
Character	0.14	0.88	0.28	0.27	0.93(r=5, s=3)
Gas sensors	0.19	0.47	0.22	0.20	0.47(r=170, s=2)
平均值	0.42	0.67	0.57	0.57	0.80

表 4 *nSeeds*=5 时各种方法的 *Precision*

数据集	D_{Eros}	中心序列	Li's first	Li's second	2DSVD
Lp1	0.88	1.00	1.00	1.00	1.00(r=7, s=3)
Lp2	0.75	0.89	0.99	0.99	1.00(r=15, s=1)
Lp4	0.66	0.91	0.65	0.70	1.00(r=15, s=6)
Lp5	0.48	0.94	0.54	0.53	0.98(r=15, s=5)
BCI	0.50	0.46	0.45	0.45	0.45(r=500, s=2)
Vowel	0.19	0.54	0.81	0.81	0.82(r=1, s=12)
Wafer	0.22	0.40	0.10	0.11	0.51(r=28, s=1)
AUSLAN	0.41	0.32	0.81	0.87	0.92(r=1, s=21)
Character	0.15	0.90	0.35	0.27	0.96(r=5, s=3)
Gas sensors	0.19	0.49	0.22	0.23	0.48(r=170, s=2)
平均值	0.44	0.69	0.59	0.60	0.81

3.3 参数对半监督分类性能的影响

本文提出的分类算法有两个参数: 一个是行-行协方差矩阵的主要特征向量个数 r, 另一个是列-列协方差矩阵的主要特征向量个数 s. 图 1、图 2 分别给出了在 AUSLAN、Vowel 数据集上, 将参数 r 固定为 1, *Precision* 随参数 s 的变化情况. 从图 1 和图 2 可以看

出, 当 s=1 时, *Precision* 最小; 随着 s 逐渐增加, 算法的 *Precision* 快速上升, 然后趋于平稳; 所以, 在算法的执行过程中, 可以选取较大的 s 值来提高分类的 *Precision*.

表 5 Wilcoxon 符号秩检验

检验量	Signedrank 值	概率 p 值
	nSeeds=1	
D_{Eros} 与 2DSVD	1	0.0039
Li's First 与 2DSVD	1	0.0156
Li's Second 与 2DSVD	1	0.0156
中心序列与 2DSVD	4	0.0234
	nSeeds=3	
D_{Eros} 与 2DSVD	1	0.0039
Li's First 与 2DSVD	1	0.0078
Li's Second 与 2DSVD	0	0.0039
中心序列与 2DSVD	0	0.0156
	nSeeds=5	
D_{Eros} 与 2DSVD	1	0.0039
Li's First 与 2DSVD	0	0.0078
Li's Second 与 2DSVD	0	0.0078
中心序列与 2DSVD	3	0.0195

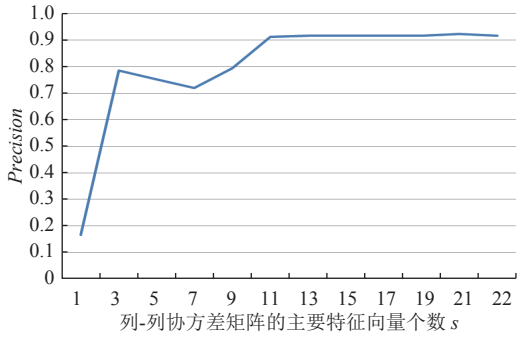

图 1 AUSLAN 数据集 *Precision* 随列-列协方差矩阵的主要特征向量个数 s 的变化

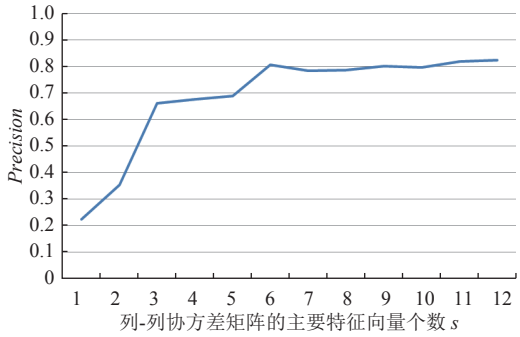

图 2 Vowel 数据集 *Precision* 随列-列协方差矩阵的主要特征向量个数 s 的变化

图 3 给出了在 AUSLAN 数据集上, 将参数 s 固定为 21, *Precision* 随参数 r 的变化情况. 图 4 给出了在 Vowel 数据集上, 将参数 s 固定为 12, *Precision* 随参数 r 的变化情况. 从图 3 和图 4 可以看出, 当参数 r 增加时, 分类的 *Precision* 趋于平稳; 所以, 在算法执行过程中, 可以选取适当的 r 值即可.

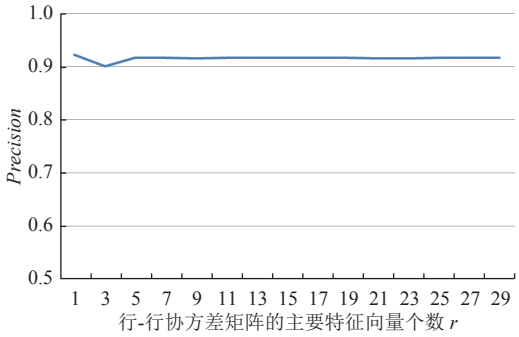

图 3 AUSLAN 数据集 *Precision* 随行-行协方差矩阵的主要特征向量个数 r 的变化

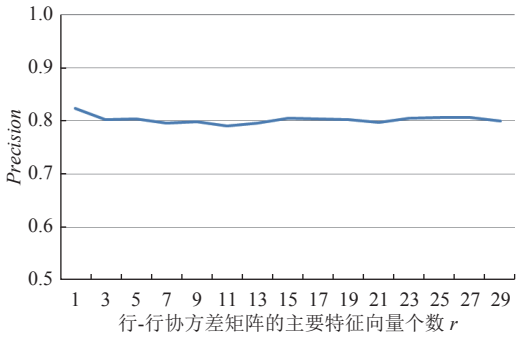

图 4 Vowel 数据集 *Precision* 随行-行协方差矩阵的主要特征向量个数 r 的变化

在本文实验中, 参数 r 和 s 的选取方法如下[2]: 首先选择一个较大的 s 值, 使得这 s 个列-列协方差矩阵的主要特征向量能够描述列-列之间总变异 (total column-column variations) 的 98% 或 99%, 其次, 让 r 值从 1 增加到 m, 其中 m 为观测值个数, 计算相对于每一个 r 值的所有训练样本的重构误差平方和, 最后根据重构误差平方和的相对变化情况选取适当的参数 r.

4 结论与展望

本文提出了一种基于 2DSVD 的 MTS 半监督分类方法, 在 10 个 MTS 数据集上对该方法进行验证, 实验结果表明, 本文提出的算法显著地好于基于一维 SVD 的 Li's First、Li's Second 方法[15,16],基于扩展 Frobenius 范数的距离 D_{Eros}[26], 以及中心序列[27]. 虽然本文建立的是一类分类器, 因此也可以很容易地修改本文提出的

算法以适应多类问题. 本文提出的算法有两个参数 r 和 s, 如何自动地选择最优的 r 和 s 值以及选取更优的分类器和停止标准值得今后进一步研究.

参考文献

1 马超红, 翁小清. 基于 PAA 的时间序列早期分类. 计算机科学, 2018, 45(2): 291–296, 317. [doi: 10.11896/j.issn.1002-137X.2018.02.050]

2 翁小清. 多变量时间序列的异常识别与分类研究[博士学位论文]. 西安: 西安交通大学, 2008.

3 Chen YP, Hu B, Keogh E, *et al*. DTW-D: Time series semi-supervised learning from a single example. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, IL, USA. 2013. 383–391.

4 Ding C, Ye JP. Two-dimensional singular value decomposition (2dsvd) for 2d maps and images. SIAM International Conference Data Mining. 2005. 32–43.

5 Weng XQ, Shen JY. Classification of multivariate time series using two-dimensional singular value decomposition. Knowledge-Based Systems, 2008, 21(7): 535–539. [doi: 10.1016/j.knosys.2008.03.014]

6 单中南, 翁小清, 马超红. 时间序列半监督分类综述. 河北省科学院学报, 2018, 35(2): 49–54.

7 单中南, 翁小清, 武天鸿. 基于 LPP 的时间序列半监督分类. 智能计算机与应用, 2019, 9(1): 6–13. [doi: 10.3969/j.issn.2095-2163.2019.01.002]

8 Wei L, Keogh E. Semi-supervised time series classification. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, USA. 2006. 748–753.

9 Ratanamahatana CA, Wanichsan D. Stopping criterion selection for efficient semi-supervised time series classification. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. Berlin, Germany. 2008. 1–14.

10 Begum N, Hu B, Rakthanmanon T, *et al*. Towards a minimum description length based stopping criterion for semi-supervised time series classification. 2013 IEEE 14th International Conference on Information Reuse & Integration. San Francisco, CA, USA. 2013. 333–340.

11 Begum N, Hu B, Rakthanmanon T, *et al*. A minimum description length technique for semi-supervised time series classification. In: Bouabana-Tebibel T, Rubin S, eds. Integration of Reusable Systems. Cham: Springer International Publishing, 2014. 171–192.

12 Vinh VT, Anh DT. Some novel improvements for MDL-based semi-supervised classification of time series. International Conference on Computational Collective Intelligence. Seoul, South Korea. 2014. 483–493.

13 Vinh VT, Anh DT. Two novel techniques to improve MDL-based semi-supervised classification of time series. In: Nguyen N, Kowalczyk R, Orłowski C, *et al*, eds. Transactions on Computational Collective Intelligence XXV. Berlin, Heidelberg: Springer, 2016. 127–147.

14 Vinh VT, Anh DT. Constraint-based MDL principle for semi-supervised classification of time series. 2015 Seventh International Conference on Knowledge and Systems Engineering. Ho Chi Minh City, Vietnam. 2016. 43–48.

15 Li CJ, Khan L, Prabhakaran B. Real-time classification of variable length multi-attribute motions. Knowledge and Information Systems, 2006, 10(2): 163–183. [doi: 10.1007/s10115-005-0223-8]

16 Li CJ, Khan L, Prabhakaran B. Feature selection for classification of variable length multiattribute motions. Multimedia Data Mining and Knowledge Discovery. New York, NY, USA. 2007. 116–137.

17 Aha DW. Feature weighting for lazy learning algorithms. Feature Extraction, Construction and Selection. Boston, MA, USA. 1998. 13–32.

18 Blankertz B, Curio G, Muller K. Classifying single trial EEG: Towards brain computer interfacing. Advances in Neural Information Processing Systems. Vancouver, BC, Canada. 2002. 157–164.

19 Schlögl A, Neuper C, Pfurtscheller G. Estimating the mutual information of an EEG-based Brain-Computer Interface. Biomedizinische Technik Biomedical Engineering, 2002, 47(1–2): 3–8.

20 UCI KDD Archive. http://kdd.ics.uci.edu/summary.data.type.html.

21 Bobski's world. http://www.cs.cmu.edu/~bobski/.

22 Williams BH. Second year PhD report extracting motion primitives from natural handwriting data. International Conference on Artificial Neural Networks. Berlin, Germany. 2006.

23 Vergara A, Vembu S, Ayhan T, *et al*. Chemical gas sensor drift compensation using classifier ensembles. Sensors and Actuators B: Chemical, 2012, 166–167: 320–329. [doi: 10.1016/j.snb.2012.01.074]

24 Rodriguez-Lujan I, Fonollosa J, Vergara A, *et al*. On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. Chemometrics and Intelligent Laboratory Systems, 2014, 130: 123–134. [doi:

10.1016/j.chemolab.2013.10.012]

25 Rodríguez JJ, Alonso CJ, Maestro JA. Support vector machines of interval-based features for time series classification. Knowledge-Based Systems, 2005, 18(4–5): 171–178. [doi: 10.1016/j.knosys.2004.10.007]

26 Yang K, Shahabi C. A PCA-based similarity measure for multivariate time series. Proceedings of the 2nd ACM International Workshop on Multimedia Databases. Washington, WA, USA. 2004. 65–74.

27 Li HL. Piecewise aggregate representations and lower-bound distance functions for multivariate time series. Physica A: Statistical Mechanics and Its Applications, 2015, 427: 10–25. [doi: 10.1016/j.physa.2015.01.063]