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

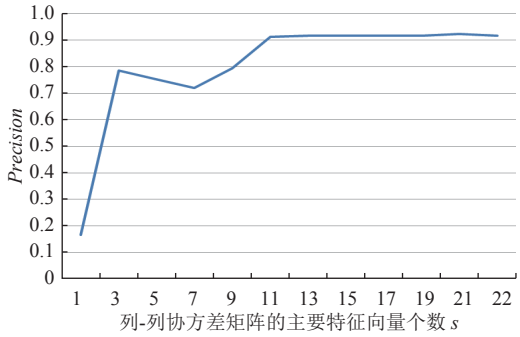

图 1 AUSLAN 数据集 *Precision* 随列-列协方差矩阵的主要特征向量个数 s 的变化

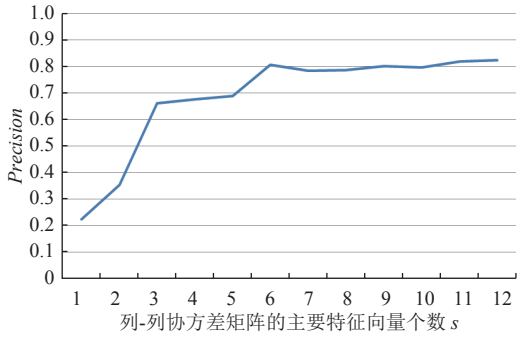

图 2 Vowel 数据集 *Precision* 随列-列协方差矩阵的主要特征向量个数 s 的变化

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.

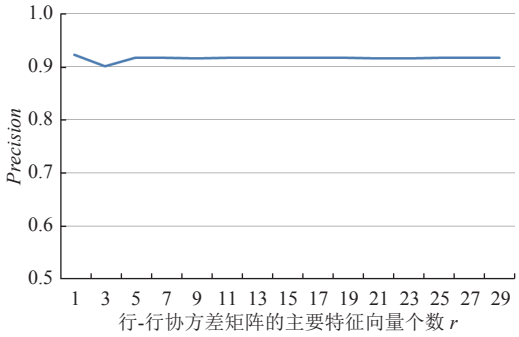

图 3 AUSLAN 数据集 *Precision* 随行-行协方差矩阵的主要特征向量个数 r 的变化

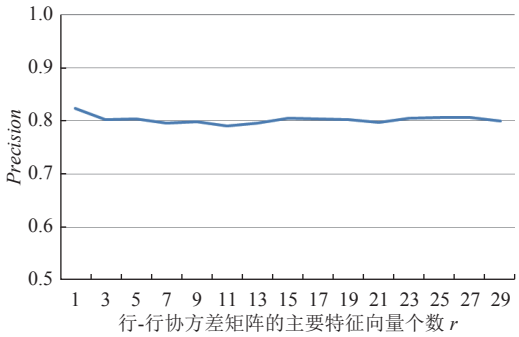

图 4 Vowel 数据集 *Precision* 随行-行协方差矩阵的主要特征向量个数 r 的变化

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]