文章编号：2095-2163(2019)01-0006-08 中图分类号：TP391 文献标志码：A

基于 LPP 的时间序列半监督分类

单中南，翁小清，武天鸿
（河北经贸大学 信息技术学院，石家庄 050061）

摘 要：在时间序列研究领域，半监督分类技术越来越受到广泛关注，绝大多数现有研究都是对原始时间序列直接进行半监督分类，一般情况下，时间序列的维数（长度）比较高，在半监督分类方法中选择合适的降维技术非常重要。本文提出了一种基于局部保持投影的时间序列半监督分类方法。该方法首先使用局部保持投影对时间序列样本进行维数约减，然后对降维后的数据进行半监督分类。在 15 个时间序列数据集的实验结果表明，该方法的分类性能显著地好于已有方法。

关键词：时间序列；局部保持映射；半监督分类；数据降维

Time series semi-supervised classification based on Locality Preserving Projections

SHAN Zhongnan, WENG Xiaoqing, WU Tianhong
(Information Technology College, Hebei University of Economics & Business, Shijiazhuang 050061, China)

【Abstract】 In the field of time series research, semi-supervised classification technology has attracted more and more attention. The existing research mainly focuses on semi-supervised classification of time series of raw data. In general, the dimension (length) of the time series is relatively high. It is very important to choose the appropriate dimensionality reduction technique in the semi-supervised classification method. This paper proposes a time-semi-supervised classification method based on Locality Preserving Projections. The method first uses the locality preserving projections to reduce the dimensionality of the time series samples, and then semi-supervised the reduced-dimensional data. The experimental results in 15 time series datasets show that the classification performance of this method is significantly better than the existing methods.

【Key words】 multivariate time series; Locality Preserving Projections; semi-supervised classification; dimensionality reduction

0 引 言

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]。已有研究发现，时间序列数据常可广泛见于金融、医学、交通等诸多领域。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数据建立分类器[2]。

目前，绝大多数已有的时间序列半监督分类（Semi-supervised classification on Time Series，SSCTS）方法，都是对原始时间序列直接进行半监督分类。由于时间序列样本的维数（即长度）随时间而不断增加，当时间序列的维数较高时，会出现维灾（curse of dimensionality）现象。为此，本文则采用了流形学习算法，在对时间序列原始数据降维的同时，还能具体考虑其在低维空间的内在结构。目前，研究指出，流形学习方法可以分为线性和非线性两种[3]。其中，非线性方法包括等距映射[4]（Isometric Mapping，IsoMap）、局部线性嵌入[5]（Locally Linear Embedding，LLE）以及拉普拉斯特征映射（Laplacian Eigenmaps，LE）[6]等方法，这些方法只能在给定的数据集上运行，对新的数据缺乏泛化能力，即都没有给出一种方法，将新的数据或对象映射到低维空间，所以这些方法不适合于半监督分类问题；线性方法包括局部保持映射[7-8]（Locality Preserving Projection，LPP）、邻域保持嵌入[9]（Neighborhood Preserving Embedding，NPE）、弹性保持映射[9]（Elastic Preserving Projections，EPP）等方法。而据分析可知，LPP 是一种线性的流形学习方法，在解决上述非线性算法存在问题的同时，还能使降维后的数据切实清晰地保持原数据的局部邻域信息。在综合了前述研究成果基础上，本文即有针对性地提出了一种基于 LPP 的时间序列半监督分类方法（LPP_SSCTS）。该方法首先使用 LPP 对时间序列

作者简介：单中南（1990-），男，硕士研究生，CCF 会员（65016G），主要研究方向：数据挖掘、机器学习；翁小清（1965-），男，博士，教授，主要研究方向：数据挖掘、机器学习；武天鸿（1993-），女，硕士研究生，主要研究方向：数据挖掘、信息检索。

收稿日期：2018-11-20

样本进行维数约减，然后对降维后的数据进行半监督分类。

本文的论述结构安排如下：首先探讨了本文研究的基础背景和相关工作；其次，提出了基于 LPP 的时间序列半监督分类算法；接下来，选取本文提出的方法与其它半监督分类方法构建仿真实验，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；最后，给出了本文研究结论。

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 $x_i(t)$ $[i = 1, 2,\dots,n;\ t = 1, 2,\dots,m]$ 表示，其中 m 是观测值的个数，n 是变量的个数[10]。当 $n = 1$ 时，称为单变量时间序列；当 $n \geqslant 2$ 时，称为多变量时间序列。本文的研究只是针对单变量时间序列。

定义 2　*P* 集合　P 为训练数据的一个集合[11]，包括所有正类标记的样本。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。随着学习的继续进行，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[11]。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　局部保持投影

局部保持映射（Locality Preserving Projection，LPP）[7-8]作为一种线性降维方法，是非线性拉普拉斯（Laplacian）特征映射的线性近似。LPP 在设计中考虑了数据的局部结构，使用 LPP 可以得到一个简单的线性变换，这个线性变换在某种特定的意义上可以最优地保持原数据集的局部邻域信息。与非线性降维方法相比，LPP 方法适用于新的样本，因此，可以将 LPP 应用于半监督分类问题。LPP 的数学定义可表述如下：

已知 R^n 中的集合 $X = \{x_1, x_2,\dots, x_k\}$，使用 LPP 方法找到变换矩阵 $\boldsymbol{A}$，将高维空间数据集 X 映射到低维空间 R^l 中的集合 $Y = \{y_1, y_2,\dots, y_k\}$，$(l << n)$，即 $y_i = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i$，使得 y_i 能够“代表”x_i。

可通过解决最小化问题来得到变换矩阵 $\boldsymbol{A}$ 中的变换变量 $\boldsymbol{a}$，对此可写作如下数学形式：

$$\min_{\boldsymbol{a}} \sum_{i,j=1}^{N} (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_j)^2 \boldsymbol{S}_{ij} \tag{1}$$

其中，$\boldsymbol{S}$ 为相似矩阵，如果 $\boldsymbol{x}_i$ 在 $\boldsymbol{x}_j$ 的 k 近邻中或 $\boldsymbol{x}_j$ 在 $\boldsymbol{x}_i$ 的 k 近邻中，$\boldsymbol{S}_{ij} = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/t)$，其它情况 $\boldsymbol{S}_{ij} = 0$，参数 t 为一个适当常数（本文设置为 1）。可以通过求解下面的广义特征值问题来使目标函数（1）达到最小，其数学公式可表示为：

$$\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{a} = \lambda \boldsymbol{XDX}^{\mathrm{T}}\boldsymbol{a} \tag{2}$$

因为 Laplacian 矩阵 $\boldsymbol{L}$ 和对角矩阵 $\boldsymbol{D}$ 都是对称的和半正定的，故 2 个矩阵 $\boldsymbol{XLX}^{\mathrm{T}}$ 和 $\boldsymbol{XDX}^{\mathrm{T}}\boldsymbol{y}$ 也是对称的和半正定的。设 $\boldsymbol{a}_0,\dots,\boldsymbol{a}_{d-1}$ 为式（2）的解，并已根据其特征值排序：$0 \leqslant \lambda_0 \leqslant \dots \leqslant \lambda_{d-1}$。则嵌入结果将分别如式（3）~（4）所示：

$$\boldsymbol{x}_i \rightarrow \boldsymbol{y}_i = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i \tag{3}$$

$$\boldsymbol{A} = (\boldsymbol{a}_0, \boldsymbol{a}_1,\dots,\boldsymbol{a}_{l-1}) \tag{4}$$

其中，$\boldsymbol{y}_i$ 为 l 维向量，$\boldsymbol{A}$ 为 $n * l$ 矩阵。

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[2]，即：基于实例、基于聚类以及基于模型的半监督分类方法。对此，可做阐释解析如下。

Chen 等人[11]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。Wei 等人[12]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 等人[13]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；Begum 等人[14-15]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。针对此问题，Vinh 等人[16-17]提供了后续改进，并增加一个后处理步骤，使分类器更加精确。Vinh 等人[18]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，必须满足约束条件 $DL(t \mid H) < DL(t)$，才能添加到正类集合。另外，Vinh 等人还定义了安全距离（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。

Nguyen 等人[19]提出了一种 PU 学习算法 LCLC（Learning from Common Local Clusters），可以从未标记的集合 U 中有效地提取正类和负类样本。LCLC

是基于聚类的方法,采用了特征选择策略,考虑了正类以及未标记实例的特征,从而使 LCLC 能够更准确地评估簇或样本之间的相似性。Nguyen 等人[20]对 LCLC 算法加以改进,提出了 En-LCLC 算法。En-LCLC 采用基于融合(ensemble)策略;通过多次执行 LCLC 算法,降低了使用单个 LCLC 预测产生的潜在偏差。

Meng 等人[21]提出了一种基于协同训练的时间序列 SSC 方法,该方法在协同训练阶段使用 HMM(hidden Markov model)和 1-NN 两种学习器(learner)。Kim[22]将模式分类问题看作是混合生成模型(generative model)的密度估计问题,将其早期提出的有判别能力的混合模型的递归估计方法,扩展到了时间序列的半监督分类;Kim[23]还提出了一种基于正则化框架(regularization framework)的半监督学习算法;将熵最小化方法、半监督支持向量机(S3VM)扩展到时间序列领域,采用 HCRF(Hidden Conditional Random Field)模型捕获时间序列数据中复杂的依赖结构。Xu 等人[24]提出了一种基于图的半学习框架,在使用 harmonic Gaussian fields 方法构造的图上,使用类标签的传播,对没有类别标签的时间序列进行分类;Nooralishahi 等人[25]基于 growing neural gas(GNG)学习框架,提出了一种在线半监督多通道(multi-channel)时间序列分类器。该方法能够处理多通道时间序列,引入了一种标签预测策略以减少误分类。

在已有的半监督分类算法中,都是直接使用时间序列原始数据进行半监督分类,由于时间序列原始数据的维数随时间而增高,存在“维灾”现象,从而影响分类性能。本文提出的基于 LPP 的半监督分类方法,对时间序列原始数据使用局部保持投影来提取高维空间数据的局部流形结构信息,在达到降维的目的同时,提高分类器的性能。

2 基于 LPP 的时间序列半监督分类算法

2.1 训练分类器

本文提出半监督分类算法主要包括 4 个步骤,各步骤内容可分述如下。

(1)对未标记的原始数据集使用主成分分析(PCA)进行预处理,目的在于去噪声处理和解决矩阵奇异性问题。

(2)构造邻接图,并对数据进行特征映射,得到降维后的数据。

(3)对降维后的数据随机选取若干个正类样本作为初始标记数据集 P。计算集合 U 中每个样本到集合 P 的欧氏距离,并将集合 U 中与集合 P 最近的样本,从集合 U 中删除,添加至集合 P。

(4)重复(3),直到满足停止标准为止。

至此将研发推得基于 LPP 的时间序列半监督分类算法的设计流程详见如下。

算法 1 LPP_SSCTS (P, U, d, k, *PCA ratio*, *nseeds*)

输入:P 表示初始训练集,包含少量已标记正类样本;U 表示未标记数据集;d 表示所降维数;k 表示近邻个数;*PCA ratio* 表示 PCA 率; *nseeds* 值表示初始标记为正类样本的个数

输出:训练好的分类器

Step1 使用 PCA 将训练集 ($P + U$) 投影到 PCA 子空间中,以达到去除噪声的目的。

Step2 用 $\boldsymbol{A}_{\mathrm{PCA}}$ 表示 PCA 的变换矩阵, $\boldsymbol{y}_i = \boldsymbol{A}_{\mathrm{PCA}}{}^{\mathrm{T}}\boldsymbol{x}_i, \boldsymbol{x}_i \in$ 训练集$(P + U)$。

Step3 在 PCA 子空间中,搜索 $\boldsymbol{y}_i$ 的 k 最近邻,构建邻接图 G。

Step4 计算相似性矩阵 $\boldsymbol{S}$,及拉普拉斯矩阵 $\boldsymbol{L}$, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$,其中 $\boldsymbol{D}$ 为原始训练集构成的对角矩阵。

Step5 令列向量 $\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{d-1}$ 为公式(2)的解,按其特征值进行排序,即:$0 \leqslant \lambda_0 \leqslant \lambda_1 \leqslant \ldots \leqslant \lambda_{d-1}$。

Step6 由列向量组成的变换矩阵 $\boldsymbol{A}_{\mathrm{LPP}}$ 与 $\boldsymbol{A}_{\mathrm{PCA}}$ 相乘得到所要求出的变换后的矩阵 $\boldsymbol{M}$。

Step7 使用 $\boldsymbol{M}$ 与原始数据 $P + U$ 相乘得到维数约减后的数据 $\boldsymbol{F}$。

Step8 从 $\boldsymbol{F}$ 中随机选取 *nseeds* 个正类样本放入集合 P。

Step9 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。

Step10 重复 Step9,直到满足停止标准为止。

在 Step10 中,本文采用 Wei 等人[12]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。LPP_SSCTS 分为 2 个阶段,对其可做解读论述如下。

(1)Step1~Step6 为数据降维阶段:设训练集中有 m 个长度为 n 的样本。Step1 中对训练集使用 PCA 进行预处理的时间复杂度为 $O(m * n^2)$, Step2~Step6 采用 LPP 对训练集进行降维可分为 2 步:

前者为 k 近邻搜索，时间复杂度为 $O((n+k)*m^2)$[7-8,10]；后者为计算特征值，若将 m 维降到 d 维，时间复杂度为 $O((n+d)*n^2)$，因此 LPP 算法的复杂度为 $O((n+k)*m^2+(n+d)*n^2))$。

(2) Step7～Step8 为训练分类器阶段，时间复杂度为 $O(m^2)$。

故而，算法的总时间复杂度为 $O((m*n^2)+(n+k)*m^2+(n+d)*n^2+m^2)$。

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[12]，阈值 r 为正类样本与其最近邻之间距离的平均值。

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对象和其它类对象。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。精确度的数学定义可表示如下：

$$precision = \frac{N_{positive}}{K} \tag{5}$$

其中，K 表示测试集中的正类样本的个数，$N_{positive}$ 表示在前 K 个最接近 P 集合的样本中，正类样本的个数。

3　实验

本节从 *Precision* 角度来评估 LPP_SSCTS 与原始方法的性能，研究采用的 15 个时间序列数据集均来自于 UCR[26] 档案库。

3.1　数据集描述

表 1 列出了 15 个时间序列数据集的主要特征，包括数据集名称、类别数、训练集样本数量、测试集样本数量以及样本长度。选用数据集来自于工业、医学、图像、生物等领域。

3.2　性能比较

将本文提出的基于 LPP 的半监督分类方法（LPP_SSCTS），与 Wei 等人[12]提出的方法（用 Wei_SSCTS 表示）的分类性能进行比较。在实验中，将数据集中类别标记为 1 的样本作为正类样本，其它类样本为负类样本。在算法 1 中，初始正类样本的个数 *nSeeds* 分别取不同值，实验重复 2 000 次，表 2～4 中给出了 *nSeeds* 分别取 1、3、5 时 2 种方法的平均 *Precision*。

表 1　数据集描述

Tab. 1　Dataset description

编号	数据集名称	类别个数	训练集样本数	测试集样本数	时间序列长度
1	Gun_Point	2	50	150	150
2	CBF	3	30	900	128
3	lighting2	2	60	60	637
4	Synthetic_Control	6	300	300	60
5	SonyAIBORobot Surface	2	20	601	70
6	Beef	5	30	30	470
7	ECG200	2	100	100	96
8	ECGFiveDays	2	23	861	136
9	Symbols	6	25	955	398
10	TwoLeadECG	2	23	1 139	82
11	FaceAll	14	560	1 690	131
12	yoga	2	300	3 000	426
13	ItalyPowerDemand	2	67	1 029	24
14	MoteStrain	2	20	1 252	84
15	Wafer	2	1 000	6 174	152

表 2～4 中的第 2～4 列分别给出了使用 Wei_SSCTS，LPP_SSCTS 的 *Precision* 以及相应参数。

分析可知，LPP_SSCTS 在 15 个数据集上分类的平均 *Precision* 高于 Wei_SSCTS 的平均 *Precision*。2 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。从表 5 中可以看到，当 *nSeeds* 为 1、3、5 时，LPP_SSCTS 与 Wei_SSCTS 的 Wilcoxon 符号秩检验的概率 p 值都小于 0.05，说明 LPP_SSCTS 的分类性能显著地好于 Wei_SSCTS。

3.3　参数对半监督分类性能的影响

本文提出的 LPP_SSCTS 算法有 3 个参数，即：*PCA ratio*、最近邻 k 的数量、以及嵌入维数 d。图 1 给出了在 Synthetic_Control 数据集上，当 *PCA ratio* = 0.66、且最近邻数 k = 10 时，*Precision* 随嵌入维数 d 的变化情况。图 2 给出了在 FaceAll 数据集上，当 *PCA ratio* = 0.86、且最近邻数 k = 6 时，*Precision* 随嵌入维数 d 的变化情况。从图 1 和图 2 可以看出，嵌入维数 d 对算法的性能有较大影响。当嵌入维数 d 比较小时，*Precision* 比较低。产生这种情况，一种可能的解释为数据集中不同的区域经过映射以后，在嵌入空间中重叠在了一起；随着嵌入维数 d 逐步增加，*Precision* 快速上升。

表 2 *nSeeds* = 1 时各种方法的 *Precision*

Tab. 2 *Precision* of various methods when *nSeeds* = 1

数据集	*nSeeds* = 1		
	Wei_SSCTS	LPP_SSCTS	参数
Gun_Point	0.61	0.63	$d = 50$, *PCA ratio* = 0.99, $k = 4$
CBF	0.34	0.35	$d = 27$, *PCA ratio* = 0.98, $k = 5$
lighting2	0.52	0.62	$d = 2$, *PCA ratio* = 0.97, $k = 5$
Synthetic_Control	0.19	0.76	$d = 3$, *PCA ratio* = 0.66, $k = 10$
SonyAIBORobot Surface	0.40	0.56	$d = 2$, *PCA ratio* = 0.96, $k = 8$
Beef	0.63	0.81	$d = 15$, *PCA ratio* = 0.99, $k = 6$
ECG200	0.78	0.78	$d = 48$, *PCA ratio* = 0.79, $k = 6$
ECGFiveDays	0.52	0.53	$d = 22$, *PCA ratio* = 0.95, $k = 4$
Symbols	0.52	0.53	$d = 24$, *PCA ratio* = 0.77, $k = 11$
TwoLeadECG	0.53	0.53	$d = 21$, *PCA ratio* = 0.96, $k = 11$
FaceAll	0.45	0.46	$d = 23$, *PCA ratio* = 0.86, $k = 6$
yoga	0.43	0.47	$d = 2$, *PCA ratio* = 0.87, $k = 11$
ItalyPowerDemand	0.61	0.64	$d = 3$, *PCA ratio* = 0.99, $k = 5$
MoteStrain	0.61	0.59	$d = 19$, *PCA ratio* = 0.70, $k = 15$
Wafer	0.88	0.90	$d = 76$, *PCA ratio* = 0.98, $k = 5$
平均值	0.53	0.61	

表 3 *nSeeds* = 3 时各种方法的 *Precision*

Tab. 3 *Precision* of various methods when *nSeeds* = 3

数据集	*nSeeds* = 3		
	原始数据	LPP_SSCTS	参数
Gun_Point	0.61	0.65	$d = 50$, *PCA ratio* = 0.99, $k = 4$
CBF	0.37	0.39	$d = 27$, *PCA ratio* = 0.98, $k = 5$
lighting2	0.52	0.64	$d = 2$, *PCA ratio* = 0.97, $k = 5$
Synthetic_Control	0.20	0.85	$d = 3$, *PCA ratio* = 0.66, $k = 10$
SonyAIBORobot Surface	0.30	0.59	$d = 2$, *PCA ratio* = 0.96, $k = 8$
Beef	0.79	0.92	$d = 15$, *PCA ratio* = 0.99, $k = 6$
ECG200	0.78	0.78	$d = 48$, *PCA ratio* = 0.79, $k = 6$
ECGFiveDays	0.53	0.57	$d = 22$, *PCA ratio* = 0.95, $k = 4$
Symbols	0.80	0.84	$d = 24$, *PCA ratio* = 0.77, $k = 11$
TwoLeadECG	0.55	0.60	$d = 21$, *PCA ratio* = 0.96, $k = 11$
FaceAll	0.64	0.69	$d = 23$, *PCA ratio* = 0.86, $k = 6$
yoga	0.44	0.47	$d = 2$, *PCA ratio* = 0.87, $k = 11$
ItalyPowerDemand	0.61	0.64	$d = 3$, *PCA ratio* = 0.99, $k = 5$
MoteStrain	0.61	0.65	$d = 19$, *PCA ratio* = 0.70, $k = 15$
Wafer	0.91	0.92	$d = 76$, *PCA ratio* = 0.98, $k = 5$
平均值	0.58	0.66	

表 4　*nSeeds*=5 时各种方法的 *Precision*

Tab. 4　*Precision* of various methods when *nSeeds*=5

数据集	*nSeeds* = 5		
	Wei_SSCTS	LPP_SSCTS	参数
Gun_Point	0.61	0.65	d = 50, *PCA ratio* = 0.99, k = 4
CBF	0.38	0.42	d = 27, *PCA ratio* = 0.98, k = 5
lighting2	0.52	0.65	d = 2, *PCA ratio* = 0.97, k = 5
Synthetic_Control	0.20	0.87	d = 3, *PCA ratio* = 0.66, k = 10
SonyAIBORobot Surface	0.30	0.60	d = 2, *PCA ratio* = 0.96, k = 8
Beef	0.83	1.00	d = 15, *PCA ratio* = 0.99, k = 6
ECG200	0.78	0.78	d = 48, *PCA ratio* = 0.79, k = 6
ECGFiveDays	0.56	0.59	d = 22, *PCA ratio* = 0.95, k = 4
Symbols	0.86	0.88	d = 24, *PCA ratio* = 0.77, k = 11
TwoLeadECG	0.57	0.64	d = 21, *PCA ratio* = 0.96, k = 11
FaceAll	0.72	0.83	d = 23, *PCA ratio* = 0.86, k = 6
yoga	0.43	0.47	d = 2, *PCA ratio* = 0.87, k = 11
ItalyPowerDemand	0.62	0.65	d = 3, *PCA ratio* = 0.99, k = 5
MoteStrain	0.63	0.70	d = 19, *PCA ratio* = 0.70, k = 15
Wafer	0.92	0.93	d = 76, *PCA ratio* = 0.98, k = 5
平均值	0.60	0.71	

表 5　Wilcoxon 符号秩检验

Tab. 5　Wilcoxon signed rank test

	nSeeds = 1		*nSeeds* = 3		*nSeeds* = 5	
	signedrank 值	概率 p 值	signedrank 值	概率 p 值	signedrank 值	概率 p 值
Wei_SSCTS 与 LPP_SSCTS	6	0.003 4	0	1.220 7e−04	0	1.220 7e−04

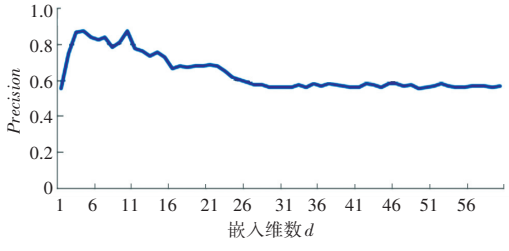

图 1　Synthetic_Control 数据集 *Precision* 随嵌入维数 *d* 的变化

Fig. 1　The *Precision* of Synthetic_Control dataset changes with the number of embedding dimension *d*

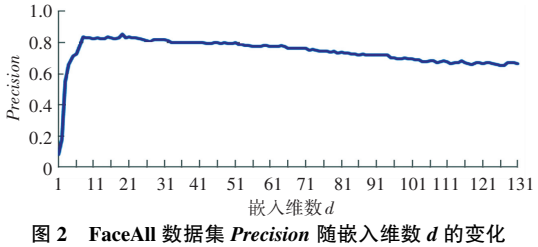

图 2　FaceAll 数据集 *Precision* 随嵌入维数 *d* 的变化

Fig. 2　The *Precision* of FaceAll dataset changes with the number of embedding dimension *d*

图 3 给出了在 Beef 数据集上，当 *PCA ratio* = 0.99、嵌入维数 d = 15 时，*Precision* 随 k 近邻个数的变化情况。图 4 给出了在 ECG200 数据集上，当 *PCA ratio* = 0.79、嵌入维数 d = 48 时，*Precision* 随 k 近邻个数的变化情况。从图 3 和图 4 中可以看出，*Precision* 在一定区域内波动，k 值对算法的性能影响相对较小。

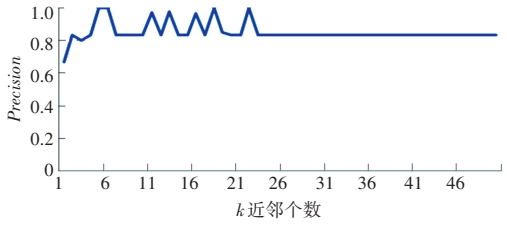

图 3　Beef 数据集 *Precision* 随 *k* 近邻个数的变化

Fig. 3　The *Precision* of Beef dataset changes with the number of nearest neighbor *k*

图 5 给出了在 Symbols 数据集上，当 k = 11、嵌入维数 d = 24 时，*Precision* 随 *PCA ratio* 的变化情况。图 6 给出了在 Synthetic_Control 数据集上，当 k =10、嵌入维数 d 为 3 时，*Precision* 随 *PCA ratio* 的变化情况。从图 5 和图 6 可以看出，*Precision* 在一

定区域内波动,*PCA ratio* 对算法的性能影响相对较小。

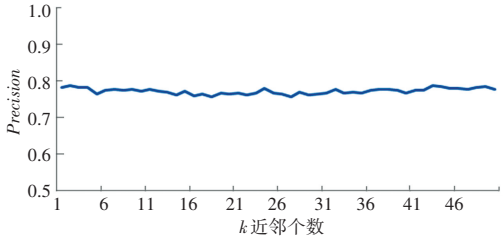

图 4 ECG200 数据集 *Precision* 随 *k* 近邻个数的变化

Fig. 4 The *Precision* of ECG200 dataset changes with the number of nearest neighbor *k*

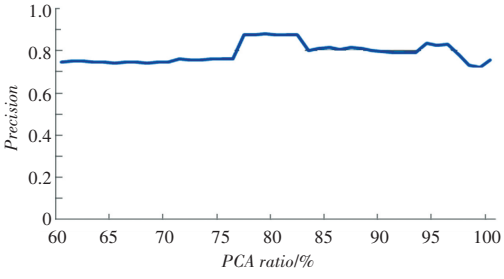

图 5 Symbols 数据集 *Precision* 随 PCA 率变化的情况

Fig. 5 The *Precision* of Symbols dataset changes with *PCA ratio*

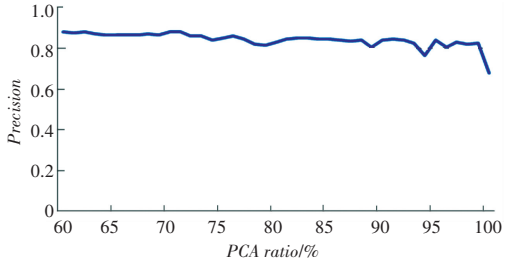

图 6 Synthetic_Control 数据集 *Precision* 随 PCA 率变化的情况

Fig. 6 The *Precision* of Synthetic_Control dataset changes with *PCA ratio*

4 结束语

本文提出了一种基于局部保持投影的时间序列半监督分类方法 LPP_SSCTS。针对不同的数据集,LPP_SSCTS 只需选择恰当的参数就可以在解决维灾和去除噪声的同时,还能使降维后的数据可以清晰地保持原数据的局部邻域信息。在 15 个时间序列数据集上的实验结果表明,本文提出的算法显著地好于 Wei_SSCTS。如何选择最优的参数以及如何将 LPP_SSCTS 应用于多变量时间序列的半监督分类,仍将亟待下一步的深入探索与研究。

参考文献

[1] 马超红, 翁小清. 时间序列早期分类综述[J]. 微型机与应用, 2016, 35(16):13-15,19.

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

[3] 罗廷金. 基于流形学习的数据降维算法研究[D]. 长沙:国防科学技术大学, 2013.

[4] TENENBAUM J B, DE SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500):2319-2323.

[5] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500):2323-2326.

[6] BELKIN M, NIYOGI P. Laplacian Eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6):1373-1396.

[7] HE Xiaofei, NIYOGI P. Locality Preserving Projections (LPP)[J]. Advances in Neural Information Processing Systems, 2002, 16(1):186-197.

[8] HE Xiaofei. Locality preserving projections[D]. Chicago, IL, USA: The University of Chicago, 2005.

[9] HE X, CAI D, YAN S, et al. Neighborhood Preserving Embedding(NPE)[C]// Tenth IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2005:1208-1213.

[10] 翁小清, 沈钧毅. 多变量时间序列的异常识别与分类研究[D]. 西安:西安交通大学,2008.

[11] CHEN Yanping, HU Bing, KEOGH E, et al. DTW-D: Time series semi-supervised learning from a single example[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, Illinois, USA: ACM, 2013:383-391.

[12] WEI Li, KEOGH E. Semi-supervised time series classification[C]// Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Philadelphia, PA, USA: ACM, 2006:748-753.

[13] RATANAMAHATANA C A, WANICHSAN D. Stopping criterion selection for efficient semi-supervised time series classification[M]// LEE R. Software engineering, artificial intelligence, networking and parallel/distributed computing. studies in computational intelligence, Berlin/ Heidelberg: Springer,2008, 149:1-14.

[14] BEGUM N, HU Bing, RAKTHANMANON T, et al. Towards a minimum description length based stopping criterion for semi-supervised time series classification[C]// 2013 IEEE 14th International Conference on Information Reuse and Integration. San Francisco, CA, USA.:IEEE, 2013:333-340.

[15] BEGUM N, HU Bing, RAKTHANMANON T, et al. A minimum description length technique for semi-supervised time series classification[M]// BOUABANA-TEBIBEL T, RUBIN S. Integration of reusable systems. Advances in Intelligent Systems and Computing, Cham: Springer, 2014,263:171-192.

[16] VINH V T, ANH D T. Some novel improvements for MDL-based semi-supervised classification of time series[M]//HWANG D, JUNG J J, NGUYEN N T. Computational collective intelligence. technologies and applications. ICCCI 2014. Lecture Notes in Computer Science, Cham: Springer, 2014, 8733: 483-493.

[17] VINH V T, ANH D T. Two novel techniques to improve MDL-based semi-supervised classification of time series[M]// NGUYEN N, KOWALCZYK R, ORLOWSKI C, ZIÓŁKOWSKI A. Transactions on computational collective intelligence XXV. lecture notes in computer science, Berlin/Heidelberg: Springer,

2016,9990:127-147.
[18] VINH V T, ANH D T. Constraint-based MDL principle for semi-supervised classification of time series [C]//2015 Seventh International Conference on Knowledge and Systems Engineering. HoChiMinh City, Vietnam :IEEE, 2015:43-48.
[19] NGUYEN M N, LI Xiaoli, NG S K. Positive unlabeled learning for time series classification [C]//IJCAI'11 Proceedings of the Twenty - Second International Joint Conference on Artificial Intelligence - Volume Two. Barcelona, Catalonia, Spain: ACM, 2011:1421-1426.
[20] NGUYEN M N, LI Xiaoli, NG S K. Ensemble based positive unlabeled learning for time series classification [M]// LEE S, PENG Z, ZHOU X, et al. Database systems for advanced applications. DASFAA 2012. lecture notes in computer science, Berlin/Heidelberg: Springer, 2012,7238:243-257.
[21] MENG Jun, WU Lixia, WANG Xiukun, et al. Granulation-based symbolic representation of time series and semi - supervised classification [J]. Computers & Mathematics with Applications, 2011, 62(9):3581-3590.
[22] KIM M. Semi - supervised recursive learning of discriminative mixture models for time - series classification [J]. International Journal of Fuzzy Logic & Intelligent Systems, 2013, 13(3):186-199.
[23] KIM M. Semi-supervised learning of hidden conditional random fields for time-series classification [J]. Neurocomputing, 2013, 119(16):339-349.
[24] ZHAO Xu, FUNAYA K. Time series analysis with graph-based semi-supervised learning[C]// IEEE International Conference on Data Science and Advanced Analytics. Paris, France: IEEE, 2015:1-6.
[25] NOORALISHAHI P, SEERA M, LOO C K. Online semi - supervised multi - channel time series classifier based on growing neural gas[J]. Neural Computing & Applications, 2016, 28(11): 3491-3505.
[26] CHEN Yanping, KEOGH E, HU Bing, et al. The UCR time series classification archive[EB/OL]. [2017 - 01 - 03]. Http://URL www.cs.ucr.edu/~eamonn/time_series_data/.

(上接第 5 页)

同时,提出该索引的基本构建方法和数据的存储格式。之后提出的结构上的 PRP 集合合并操作、相同的 unipath 上的种子合并方法和 unipath 上的相同序列合并方法,表明索引可以极大减少用于 extension 过程计算的候选种子的数量。

本文分别在模拟数据集和真实数据集上测试索引结构和基于 BWT 的索引结构的 seeding 效果。通过比较分析 seeding 过程中各步骤中的 seeds 数量情况,发现 de Bruijn 图结构能减少更多的候选 seeds,同时保持更多的有效 seeds 比率。基于 de Bruijn 图思想构建基因组索引可以作为研究精准高效的基因组序列映射算法的基础,同时为实现多基因组上的序列比对提供研究思路。

参考文献

[1] DIDELOT X, BOWDEN R J, WILSON D J. et al. Transforming clinical microbiology with bacterial genome sequencing[J]. Nature Reviews Genetics, 2012,13(9): 601-612.
[2] FRICKE W F, RASKO D A. Bacterial genome sequencing in the clinic: Bioinformatic challenges and solutions[J]. Nature Reviews Genetics, 2014, 15(1): 49-55.
[3] BREITWIESER F P, PARDO C A, SALZBERG S L, et al. Re-analysis of metagenomic sequences from acute flaccid myelitis patients reveals alternatives to enterovirus D68 infection [J]. F1000Research, 2015, 4(1): 180.
[4] SCHNEEBERGER K, HAGMANN J, DSSOWSKI S, et al. Simultaneous alignment of short reads against multiple genomes [J]. Genome Biology, 2009, 10(9): R98.
[5] HUANG Lin, POPIC V, BALZOGLOU S, et al. Short read alignment with populations of genomes[J]. Bioinformatics, 2013, 29(13): 361-370.
[6] SIREN J, VÄLIMÄKI N, MÄKINEN V, et al. Indexing graphs for path queries with applications in genome research[J]. IEEE/ACM Transactions on Computational Biology Bioinformatics, 2014, 11(2): 375-388.
[7] VALENZUELA D, VÄLIMÄKI N, PITKÄNEN E, et al. On enhancing variation detection through pan-genome indexing[EB/OL]. [2015-07-26]. https://doi.org/10.1101/021444.
[8] DILTHEY A, COX C, IQBAL Z, et al. Improved genome inference in the MHC using a population reference graph[J]. Nat Genet, 2015, 47(6):682 - 688.
[9] LI Heng, DURBIN R. Fast and accurate long-read alignment with Burrows-Wheeler transform[J]. Bioinformatics, 2010, 26(5): 589-595.
[10] LANGMEAD B, TRAPNELL C, POP M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. Genome Biology,2009, 10:R25.
[11] MU J C, JIANG Hui, KIANI A, et al. Fast and accurate read alignment for resequencing[J]. Bioinformatics, 2012, 28(18): 2366-2373.