



(21)申请号 201610671152.3

(22)申请日 2016.08.15

(71)申请人 清华大学

地址 100084 北京市海淀区清华园1号

(72)发明人 孙富春 黄文炳 曹乐乐 杨豪琳

(74)专利代理机构 北京清亦华知识产权代理事务
所(普通合伙) 11201

代理人 廖元秋

(51)Int.Cl.

G06K 9/62(2006.01)

G06K 9/46(2006.01)

权利要求书3页 说明书5页

(54)发明名称

一种对时间序列进行建模并识别的方法

(57)摘要

本发明涉及一种对时间序列进行建模并识别的方法,属于机器学习领域。该方法包括将传感器采集的数据按采集时间先后顺序排列而成的动态数据构成L条时间序列,随机选择L条时间序列中的60%-80%的N条时间序列作为训练集,剩下的时间序列作为测试集;对训练集中的每一条时间序列采用线性动态系统模型进行建模,并训练集每一条时间序列的特征表示;随机从训练集中抽取J条时间序列作为字典组成字典集,从得到的训练集每一条训练时间序列的特征表示,学习出字典集的每一条时间序列的最优特征表示,同时计算每一条训练时间序列在字典集下的编码系数;用训练集的编码系数训练支持向量机模型,实现对时间序列的识别。本方法大幅降低了数据表示的复杂度,同时显著提高识别精度。

1. 一种对时间序列进行建模并识别的方法,其特征在于,该方法包括以下步骤:

1) 将传感器采集的数据按采集时间先后顺序排列而成的动态数据构成L条时间序列,随机选择L条时间序列中的60%-80%的N条时间序列作为训练集,剩下的时间序列作为测试集;

2) 对训练集中的每一条时间序列采用线性动态系统模型进行建模:

$$\begin{cases} x(t+1) = Ax(t) + v(t) \\ y(t) = Cx(t) + w(t) \end{cases} \quad (1)$$

其中,已知变量为:时间序列 $Y=(y(0), y(1), \dots, y(t), \dots, y(\tau))$ (τ 为正整数,表示时间序列的时间长度);未知变量是:序列状态 $X=(x(0), x(1), \dots, x(t), \dots, x(\tau))$,描述空间纹理的测量矩阵C,描述动态变化的转移矩阵A,模型的噪声项 $v(t)$ 和 $w(t)$;根据已知变量Y求解未知变量X,C,A,得到训练集每一条时间序列的特征表示;

3) 随机从训练集中抽取J条时间序列作为字典组成字典集,由步骤2)得到的训练集每一条训练时间序列的特征表示,学习出字典集的每一条时间序列的最优特征表示,同时计算得到每一条训练时间序列在字典集下的编码系数;

4) 用步骤3)得到的训练集的编码系数 $\alpha_1, \alpha_2, \dots, \alpha_N$ 训练支持向量机模型,实现对时间序列的识别。

2. 如权利要求1所述方法,其特征在于,所述步骤2)计算训练集所有时间序列的特征表示 $R_1, R_2, \dots, R_N$,其中每一个特征表示的具体计算过程包括:

2-1) 对时间序列Y进行矩阵分解,求解 $Y=USV^T$,即对Y进行奇异值分解,U为左奇异方向矩阵,S为奇异值矩阵,V为右奇异方向矩阵,奇异值分解可通过MATLAB软件内嵌的数值工具包求解,得到测量矩阵 $C=U$ 和序列状态 $X=SV^T$,其中 V^T 表示对矩阵V进行转置;

2-2) 设临时矩阵变量 X_0 和 X_1 ,令 $X_0=(x(0), \dots, x(t), \dots, x(\tau-1))$, $X_1=(x(1), \dots, x(t), \dots, x(\tau))$;

2-3) 计算出最优转移矩阵A,使得公式(1)尽可能拟合时间序列在时间维度上的变化规律,通过最小化 $\|X_1 - AX_0\|_F^2$,得到最优转移矩阵 $A = X_1 X_0^T (X_0 X_0^T)^{-1}$,其中 $\|\cdot\|_F^2$ 计算输入矩阵的Frobenius范数,即计算矩阵所有元素的平方和;

2-4) 对步骤2-3)得到的最优转移矩阵A进行平滑化处理: $A = U_1 S_1 V_1^T$,用 $S_1^* = 2 \frac{1}{1 + e^{-a S_1}} - 1$,其中 $e^{-a S_1}$ 为自然指数函数,a为规整因子替换矩阵 S_1 ,得到新的转移矩阵 $A = U_1 S_1^* V_1^T$;

2-5) 分别利用步骤2-1)和2-4)求解得到的测量矩阵C和最优转移矩阵A,得到观测矩阵 $O=[C; CA; CA^2; \dots; CA^k]$,其中,k为正整数, $k=1, 2, \dots, k$;

2-6) 对步骤2-5)得到的观测矩阵正交化,求解 $O^T O = U_2 S_2 V_2^T$,得到新的观测矩阵 $W = O S_2^{-1/2} U_2$,使得观测矩阵正交化,数值计算更稳定;

2-7) 计算 $R = W W^T$ 作为时间序列Y的特征表示。

3. 如权利要求1所述方法,其特征在于,所述步骤3)具体包括:

3-1) 已知变量:记训练集中第i条时间序列的特征表示为 R_i ;需要求解变量:字典集的每条时间序列的观测矩阵为 $W_1, W_2, \dots, W_J$,J为字典集时间序列的条数;字典集中每一条时间序

列的特征表示为 $D_1, D_2, \dots, D_J$; 每条训练时间序列在字典集下的编码系数为 $\alpha_i = (\alpha_{1,i}, \alpha_{2,i}, \dots, \alpha_{J,i})$, i 为遍历训练集任意一时间序列, 其中, $\alpha_{j,i}$ 表示该训练时间序列在字典集的第 j 条时间序列下的编码系数;

3-2)字典特征表示的初始化:随机从训练集中抽取 J 条时间序列作为字典组成字典集, 即将所随机选取的训练时间序列的观测矩阵和特征表示作为字典时间序列的观测矩阵和特征表示的初始值;

3-3)对步骤3-2)所得的字典特征表示的初始值进行迭代优化, 同时计算每一条训练时间序列在字典集下的编码系数, 优化目标表达式如式(2)以最小化字典集和编码系数对训练集的重构误差并保证编码系数的稀疏性

$$D_1, \dots, D_J, \alpha_1, \dots, \alpha_N = \min \sum_i^N \left(\left\| R_i - \alpha_{j,i} \sum_j^J D_j \right\|_F^2 + \lambda \sum_j^J |\alpha_{j,i}| \right) \quad (2)$$

其中 $\lambda=0.1$ 为惩罚系数, λ 的值根据识别精度要求选定, 此处为经验值, $\|\bullet\|_F^2$ 求解矩阵的 Frobenius 范数, $|\bullet|$ 计算绝对值。

4. 如权利要求1所述方法, 其特征在于, 所述步骤3)中公式(2)求解具体过程如下:

3-3-1)利用步骤3-2)字典集的特征表示的初始值, 对训练集的每一条时间序列进行稀疏编码, 即求解式(3)的目标函数表达式, 得到训练集每一条时间序列在该字典集的编码系数:

$$\alpha_i = \min \left\| R_i - \alpha_j \sum_j^J D_j \right\|_F^2 + \lambda \sum_j^J |\alpha_{j,i}| \quad (3)$$

$i \in N$ 遍历训练集的所有时间序列, 公式(3)通过MATLAB软件内嵌的数值工具包求解;

3-3-2)得到训练集每一条时间序列的编码系数后, 计算公式(2)对应于每个字典时间序列的特征表示 D_r 的导数 ∇D_r , r 为遍历字典集任意一时间序列:

$$\nabla D_r = \sum_i^N \sum_j^J \alpha_{r,i} D_j W_r - R_i W_r \quad (4)$$

3-3-3)由步骤3-3-2)所得 ∇D_r , 更新每个字典的观测矩阵 W_r :

$$M = \left(I + \frac{1}{2} (\nabla D_r W_r^T - W_r \nabla D_r^T) \right)^{-1} \left(I - \frac{1}{2} (\nabla D_r W_r^T - W_r \nabla D_r^T) \right) W_r \quad (5)$$

$$W_r = M \quad (6)$$

其中, I 为单位矩阵;

3-3-4)由步骤3-3-3)所得 W_r , 更新每个字典的特征表示 $D_r = W_r W_r^T$;

3-3-5)重复步骤3-3-1)至3-3-4)直到公式(2)中目标函数值的变化小于设定的阈值, 得到字典集的最优特征表示 $D_1, D_2, \dots, D_J$, 以及每一条训练时间序列在最优字典集下的编码系数 $\alpha_1, \alpha_2, \dots, \alpha_N$ 。

5. 如权利要求1所述方法, 其特征在于, 所述步骤4)具体包括:

4-1)设支持向量机的模型参数为 w, b , 则该模型参数通过式(7)表示的目标函数求得:

$$\begin{aligned} \min_{w, b} & \|w\| \\ \text{s.t.} & I_i(w^T \alpha_i - b) \geq 1, i = 1, 2, \dots, N \end{aligned} \quad (7)$$

其中, $l_i \in (-1, 1)$ 表示第 i 个训练时间序列的类别, -1 代表负样本, 1 代表正样本, a_i 是训练集第 i 条时间序列的编码系数, 公式(7)的参数 w, b 通过MATLAB软件内嵌的数值工具包求解;

4-2) 对于测试集的每一条时间序列 R_{test} , 通过4-1)的方法求解出其在最优特征表示的字典集下的编码系数, 设为 a_{test} ;

4-3) 计算 $w^T a_{\text{test}} - b$ 的值, 若大于0则为正样本, 若小于零则为负样本, 从而实现了测试时间序列的类别判断, 完成了时间序列的识别。

一种对时间序列进行建模并识别的方法

技术领域

[0001] 本发明涉及一种对时间序列进行建模并识别的方法,属于机器学习领域。

背景技术

[0002] 近年来,随着摄像机和力传感器等不同传感器感知技术的发展,数据产生和采集的速度越来越快,数据存储量也越来越大,且大部分数据是以时间序列的形态进行传输和存储。所谓时间序列,是指传感器采集的数据按采集时间先后顺序排列而成的动态数据。面对海量的时间序列数据,单纯的人力已经不能有效对其进行分析并提取有用的特征信息。因此,如何设计高效的数据分析算法,对不同传感器采集到的时间序列进行有机提炼,得到富含信息量的有效表征,以及利用有效表征实现目标识别任务,成为大数据环境下开展动态数据挖掘的关键问题。

[0003] 一方面,传统的目标识别任务大多数是基于静态数据,如通过分析单张图片识别或者检测图片中的目标、基于单个时间节点的触觉力分布来识别机器人所抓取的物体。然而,通常情况下,传感器采集到的是具有特定时间长度的数据流。这需要通过分析一段视频(而不是单张图片)来识别视觉目标、通过分析整个抓取过程的触觉力的变化过程(而不是单个时间节点的力)来识别机器人的抓取目标。另一方面,针对动态数据,当前主流的方法是把输入数据截取(或采样)为时间维度等长的数据块,把这些数据块按时空展开作为高维的静态数据,然后采用传统基于静态数据的方法进行建模和识别。然而,该类方法存在两个主要的问题。首先,不同的识别目标对应于不同的物理过程,其动态数据的自然时间长度通常是不相等的,把所有数据进行等长处理会导致某些数据失真;其次,如果截取的时间长度过短,会丢失大量有用的信息,如果截取的时间长度过长,则导致数据块的维度过大,大幅降低后期识别算法的运行效率。因此,时间序列的识别亟需一种有效的建模工具来提取其有用且精炼的信息,提高识别精度并保证较快的运行效率。

[0004] 以下为几个相关技术。

[0005] 线性动态系统模型是对时间序列的每一帧数据以及相邻两帧数据的转化关系进行建模,如公式(1)所示:

$$\begin{cases} x(t+1) = Ax(t) + v(t) \\ y(t) = Cx(t) + w(t) \end{cases} \quad (1)$$

[0007] 其中,已知变量为:时间序列 $Y=(y(0), y(1), \dots, y(t), \dots, y(\tau))$ (τ 为正整数,表示时间序列的时间长度);未知变量是:序列状态 $X=(x(0), x(1), \dots, x(t), \dots, x(\tau))$,测量矩阵 C ,转移矩阵 A ,模型的噪声项 $v(t)$ 和 $w(t)$ 。

[0008] 求解式(1)表示的线性动态系统模型分别得到序列状态 X 、测量矩阵 C 和转移矩阵 A 。测量矩阵描述了空间纹理,转移矩阵描述了动态变化。

[0009] 在机器学习领域中,实现目标识别需把数据分成两部分:训练集和测试集。训练集利用时间序列及其相应的类别标签训练识别模型;测试集用来测试识别模型的性能。常用的识别模型为线性支持向量机,已被广泛应用于图像、语音和文字等静态数据的目标识别

任务。

发明内容

[0010] 本发明的目的是为克服已有技术的不足之处,提出了一种对时间序列进行建模并识别的方法,该方法利用线性动态系统模型对时间序列进行建模得到数据的特征表示,继而引入稀疏编码和字典学习技术对所得的特征进一步编码,快速有效地实现基于时间序列的目标识别。

[0011] 本发明提出的一种对时间序列进行建模并识别的方法,其特征在于,该方法包括以下步骤:

[0012] 1)将传感器采集的数据按采集时间先后顺序排列而成的动态数据构成L条时间序列,随机选择L条时间序列中的60%-80%的N条时间序列作为训练集,剩下的时间序列作为测试集;

[0013] 2)对训练集中的每一条时间序列采用线性动态系统模型进行建模:

$$[0014] \quad \begin{cases} x(t+1) = Ax(t) + v(t) \\ y(t) = Cx(t) + w(t) \end{cases} \quad (1)$$

[0015] 其中,已知变量为:时间序列 $Y=(y(0), y(1), \dots, y(t), \dots, y(\tau))$ (τ 为正整数,表示时间序列的时间长度);未知变量是:序列状态 $X=(x(0), x(1), \dots, x(t), \dots, x(\tau))$,描述空间纹理的测量矩阵C,描述动态变化的转移矩阵A,模型的噪声项 $v(t)$ 和 $w(t)$;根据已知变量Y求解未知变量X,C,A,得到训练集每一条时间序列的特征表示;

[0016] 3)随机从训练集中抽取J条时间序列作为字典组成字典集,由步骤2)得到的训练集每一条训练时间序列的特征表示,学习出字典集的每一条时间序列的最优特征表示,同时计算得到每一条训练时间序列在字典集下的编码系数;

[0017] 4)用步骤3)得到的训练集的编码系数 $\alpha_1, \alpha_2, \dots, \alpha_N$ 训练支持向量机模型,实现对时间序列的识别。

[0018] 本发明的特点及有益效果:

[0019] 本发明通过对时间序列建模,并利用模型的参数作为时间序列的初始特征表示,与传统基于截取时空数据块的方法相比,大幅降低了数据表示的复杂度。同时,通过稀疏编码和字典学习对初始特征表示进一步提炼,得到更加精炼的表征,显著提高识别精度。

具体实施方式

[0020] 本发明提出了一种对时间序列进行建模并识别的方法,包括以下步骤:

[0021] 1)将传感器采集的数据按采集时间先后顺序排列而成的动态数据构成L条(例如200条)时间序列,随机选择60%-80%的时间序列(设为N条)作为训练集,剩下的时间序列作为测试集(其中,训练集占全体时间序列的百分比越高,识别精度越高,可根据具体操作精度要求选定);

[0022] 2)对训练集中的每一条时间序列采用线性动态系统模型进行建模:

$$[0023] \quad \begin{cases} x(t+1) = Ax(t) + v(t) \\ y(t) = Cx(t) + w(t) \end{cases} \quad (1)$$

[0024] 其中,已知变量为:时间序列 $Y=(y(0),y(1),\cdots,y(t),\cdots,y(\tau))$ (τ 为正整数,表示时间序列的时间长度);未知变量是:序列状态 $X=(x(0),x(1),\cdots,x(t),\cdots,x(\tau))$,测量矩阵 C (描述空间纹理),转移矩阵 A (描述动态变化),模型的噪声项 $v(t)$ 和 $w(t)$;根据已知变量 Y 求解未知变量 X,C,A , (根据 Y,X,C,A ,可求得模型的噪声项 $v(t)$ 和 $w(t)$,本发明方法不涉及该两项,不再做具体阐述)得到训练集每一条时间序列的特征表示;具体包括:

[0025] 2-1)对时间序列 Y 进行矩阵分解,求解 $Y=USV^T$,即对 Y 进行奇异值分解, U 为左奇异方向矩阵, S 为奇异值矩阵, V 为右奇异方向矩阵,奇异值分解可通过MATLAB软件内嵌的数值工具包求解,得到测量矩阵 $C=U$ 和序列状态 $X=SV^T$,其中 V^T 表示对矩阵 V 进行转置;

[0026] 2-2)设临时矩阵变量 X_0 和 X_1 ,令 $X_0=(x(0),\cdots,x(t),\cdots,x(\tau-1))$, $X_1=(x(1),\cdots,x(t),\cdots,x(\tau))$;

[0027] 2-3)计算出最优转移矩阵 A ,使得公式(1)尽可能拟合时间序列在时间维度上的变化规律,即通过最小化 $\|X_1-AX_0\|_F^2$,得到最优转移矩阵 $A=X_1X_0^T(X_0X_0^T)^{-1}$,其中 $\|\cdot\|_F^2$ 计算输入矩阵的Frobenius范数,即计算矩阵所有元素的平方和;

[0028] 2-4)对步骤2-3)得到的最优转移矩阵 A 进行平滑化处理(避免计算中出现数值不稳定的问题),即对矩阵 A 进行奇异值分解, $A=U_1S_1V_1^T$ (对 A 奇异值分解,定义同2-1),用 $S_1'=2\frac{1}{1+e^{-aS_1}}-1$ (其中 e^{-aS_1} 为自然指数函数, a 为规整因子,经验取值2.5)替换矩阵 S_1 ,得到新的转移矩阵 $A=U_1S_1'V_1^T$;

[0029] 2-5)分别利用步骤2-1)和2-4)求解得到的测量矩阵 C 和最优转移矩阵 A ,得到观测矩阵 $O=[C;CA;CA^2;\cdots;CA^k]$, (其中, k 为正整数, $k=1,2,\cdots,k;k$ 根据经验取值,本实施例取值为5);

[0030] 2-6)对步骤2-5)得到的观测矩阵正交化,求解 $O^TO=U_2S_2V_2^T$ (对 O^TO 进行奇异值分解,定义同2-1)),得到新的观测矩阵 $W=OS_2^{-1/2}U_2$ (目的是使得观测矩阵正交化,使得数值计算更稳定);

[0031] 2-7)令 $R=WW^T$ 作为时间序列 Y 的特征表示,由 W 计算出 WW^T ;由此可计算出训练集每一条时间序列的特征表示,记为 $R_1,R_2,\cdots,R_N$;

[0032] 3)随机从训练集中抽取 J 条时间序列作为字典组成字典集,由步骤2)得到的训练集每一条训练时间序列的特征表示,学习出字典集的每一条时间序列的最优特征表示,同时计算得到每一条训练时间序列在字典集下的编码系数;具体包括:

[0033] 3-1)已知变量:记训练集中第 i 条时间序列的特征表示为 R_i ,需要求解变量:字典集的每条时间序列的观测矩阵为 $W_1,W_2,\cdots,W_J$ (J 为字典集时间序列的条数,按经验可为训练集时间序列条数的一半,观测矩阵定义见2-6));字典集中每条时间序列的特征表示为 $D_1,D_2,\cdots,D_J$;训练集每条时间序列在字典集下的编码系数为 $a_i=(a_{1,i},a_{2,i},\cdots,a_{J,i})$ (i 为遍历训练集所有时间序列中任意一时间序列),其中, $a_{j,i}$ 表示该训练时间序列在字典集的第 j 条时间序列下的编码系数;

[0034] 3-2)字典特征表示的初始化:随机从训练集中抽取 J 条时间序列作为字典组成字典集,即将所随机选取的训练时间序列的观测矩阵和特征表示作为字典时间序列的观测矩阵和特征表示的初始值;

[0035] 3-3)对3-2)所得的字典特征表示的初始值进行迭代优化,优化目标表达式如式

(2)(目的是最小化字典集和编码系数对训练集的重构误差并保证编码系数的稀疏性),同时计算每一条训练时间序列在字典集下的编码系数:

$$[0036] \quad D_1, \dots, D_J, \alpha_1, \dots, \alpha_N = \min \sum_i^N \left(\left\| R_i - \alpha_{j,i} \sum_j^J D_j \right\|_F^2 + \lambda \sum_j^J |\alpha_{j,i}| \right) \quad (2)$$

[0037] 其中 $\lambda=0.1$ 为惩罚系数, λ 的值根据识别精度要求选定,此处为经验值, $\|\bullet\|_F^2$ 求解矩阵的Frobenius范数, $|\bullet|$ 计算绝对值;公式(2)求解具体过程如下:

[0038] 3-3-1)利用步骤3-2)字典集的特征表示的初始值,对训练集的每一条时间序列进行稀疏编码,即求解式(3)的目标函数表达式,得到训练集每一条时间序列在该字典集的编码系数 α_i :

$$[0039] \quad \alpha_i = \min \left\| R_i - \alpha_j \sum_j^J D_j \right\|_F^2 + \lambda \sum_j^J |\alpha_{j,i}| \quad (3)$$

[0040] 其中, $i \in N$ 遍历训练集的所有时间序列,公式(3)通过MATLAB软件内嵌的数值工具包求解;

[0041] 3-3-2)得到训练集每一条时间序列的编码系数后,计算公式(2)对应于每个字典时间序列的特征表示 D_r (r 为遍历字典集的任一时间序列)的导数 ∇D_r :

$$[0042] \quad \nabla D_r = \sum_i^N \sum_j^J \alpha_{r,j} D_j W_r - R_i W_r \quad (4)$$

[0043] 3-3-3)由步骤3-3-2)所得 ∇D_r ,更新每个字典的观测矩阵 W_r :

$$[0044] \quad M = (I + \frac{1}{2} (\nabla D_r W_r^T - W_r \nabla D_r^T))^{-1} (I - \frac{1}{2} (\nabla D_r W_r^T - W_r \nabla D_r^T)) W_r \quad (5)$$

$$[0045] \quad W_r = M \quad (6)$$

[0046] 其中, I 为单位矩阵;

[0047] 3-3-4)由步骤3-3-3)所得 W_r ,更新每个字典的特征表示 $D_r = W_r W_r^T$;

[0048] 3-3-5)重复步骤3-3-1)至3-3-4)直到公式(2)中目标函数值的变化小于设定的阈值(阈值越小精度越高,但是需要的迭代次数就越多,按经验可取0.01),得到了字典集的最优特征表示 $D_1, D_2, \dots, D_J$,以及每一条训练时间序列在最优字典集下的编码系数 $\alpha_1, \alpha_2, \dots, \alpha_N$;

[0049] 4)用步骤3)得到的训练集的编码系数 $\alpha_1, \alpha_2, \dots, \alpha_N$,训练支持向量机模型,实现对时间序列的识别;本实施所有时间序列的类别有两类:正样本和负样本;正样本即感兴趣的目标物,如人脸、行人等,负样本是除目标物外的类别,如背景等;该步骤具体包括:

[0050] 4-1)设支持向量机的模型参数为 w, b ,则该模型参数通过以下式(7)表示的目标函数求得:

$$[0051] \quad \min_{w, b} \|w\| \quad (7)$$

$$s.t. \quad l_i(w^T \alpha_i - b) \geq 1, i = 1, 2, \dots, N$$

[0052] 其中, $l_i \in (-1, 1)$ 表示第 i 个训练时间序列的类别(-1 代表负样本, 1 代表正样本), α_i 是第 i 个训练时间序列的编码系数,公式(7)的参数 w, b 可通过MATLAB软件内嵌的数值工具包求解;

[0053] 4-2)对于测试集的每一条时间序列 R_{test} ,通过3-3-1)的方法求解出其在最优特征表示的字典集下的编码系数,设为 α_{test} ;

[0054] 4-3)计算 $w^T \alpha_{\text{test}} - b$ 的值,若大于0则为正样本,若小于零则为负样本。从而实现了测试时间序列的类别判断,完成了时间序列的识别。