第49卷第7期 中国科学技术大学学报 Vol. 49, No. 7
2019年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2019

文章编号:0253-2778(2019)07-0524-09

一种基于可伸缩模式的潜在语义挖掘方法

邱 镇[1],王琪媛[2],刘 迪[1],孟洪民[1]

(1.国网信息通信产业集团有限公司,北京 102211;2.国网(北京)节能设计研究院有限公司,北京 100052)

摘要:大数据反映了人们的生活习惯、社会规律以及自然规律.数据流作为大数据最重要的表现形式之一,应用的范围非常广泛.在实际的数据流应用领域中,连续数据点组成的波段在宏观层次上展示了丰富的语义,因此以模式(波段)为粒度来表达数据流显得尤为重要.为此基于SP-tree挖掘的可伸缩模式,提出了Pattern2vec的方法,将可伸缩模式向量化,利用向量来发现数据流上潜在的隐含语义,完成分类工作.在医疗和电力数据开展实验,实验结果表明,Pattern2vec相比其他对比方法,具有更好的分类表现.

关键词:大数据;可伸缩模式;向量化;隐含语义;分类

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.07.002

引用格式:邱镇,王琪媛,刘迪,等.一种基于可伸缩模式的潜在语义挖掘方法[J].中国科学技术大学学报,2019,49(7):524-532.

QIU Zhen, WANG Qiyuan, LIU Di, et al. A novel method for mining latent events based on scalable patterns[J]. Journal of University of Science and Technology of China, 2019, 49(7): 524-532.

A novel method for mining latent events based on scalable patterns

QIU Zhen[1], WANG Qiyuan[2], LIU Di[1], MENG Hongmin[1]

(1. *State Grid Information and Telecommunication Group Co., Ltd, Beijing* 102211, *China*;
2. *State Grid (Beijing) Energy Conservation Design and Research Institute Co., Ltd., Beijing* 100052, *China*)

Abstract: Big data reflect the people's living habits, social and natural laws. Data stream, one of the most important forms of manifestation, has a wide range of applications. In the field of practical application of data stream, the waveband consisted of continuous data point can show the abundant semantics. Therefore, it's significant to take the pattern (waveband) as the granularity and expressive form of data stream. A Pattern2vec method is proposed based on the scalable patterns mined by SP-tree, mining out the scalable patterns in data streams and then establishing the vector space for realizing the latent semantics of the scalable patterns. With the medical and electrical data sets, the experiment results show better classification performance of the proposed method.

Key words: big data; scalable pattern; vectorization; latent semantics; classification

0 引言

数据流作为当前大数据的主要形式之一,应用的范围非常广泛,包括医疗监护、生物信息、智能电网、电信业务、金融分析、天气预测等.在实际的数据流应用领域中,连续数据点组成的波段往往隐含了

收稿日期:2018-09-26;**修回日期**:2018-12-04

基金项目:国家电网科技项目(52110418002W)资助.

作者简介:邱镇(通信作者),男,1991年生,博士/工程师.研究领域:大数据、人工智能. E-mail: qiuzhen0208@126.com

潜在语义，具有较高的领域价值. 比如，在心电监护领域，随着时间的推移，不同病患的心电监护数据会呈现出不同的心电波形图（electrocardiograph，ECG）[1]. 这些波形的任何改变都可能暗示着潜在的病理[2-3]，比如波形的增加或缺失可能表明一些病理症状[4-5]. 医生可以根据心电监护数据波形变化和特点，快速、精准地为病患进行疾病诊断. 这从侧面说明了心电监护数据整体上隐含了医学语义. 文献[6]从数据流上提取了可伸缩模式，比如图 1 中可伸缩模式 PQRSTTT 和 PQRSSTTSTTSTT，虽然它们在心电监护数据上表现不同，但是它们语义相似，均表示心肌梗死的超急性损伤期现象.

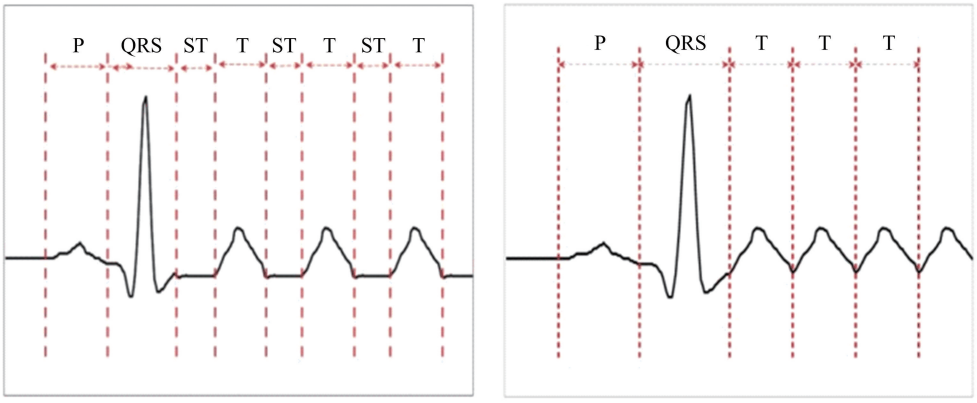

图 1　超急性损伤期的心电波形

Fig. 1　Electrocardiographic waveform in hyperacute injury stage

除医疗领域之外，在很多工业领域的相关业务分析中，数据段上发生的变化往往比单点变化更具有领域价值，比如电网数据领域，在某城市居民用电需求分析应用中，对图 2 所示的标准日用电需求曲线进行分析，可以得到日用电负荷变化规律，即从后夜 3 点左右的极低用电负荷开始呈上升趋势，至前夜 20 点左右达最高峰值. 因为这时晚间照明、广告、装饰用灯和文化活动比较集中，而中午由于就餐、午休导致的停工停产会使曲线有短暂下降态势，继而继续升高. 这就形成了几个比较突出的波峰、波谷模式，如前夜时段模式、后夜时段模式等. 在实际日用电负荷曲线上，可能会出现不同于标准曲线的情况，如某日用电负荷曲线变为两个类似后夜时段的模式，如果这种情况连续出现，则预示着用户用电行为出现异常，需要供电公司进行排查，以防出现窃电等引起安全隐患的用电情况.

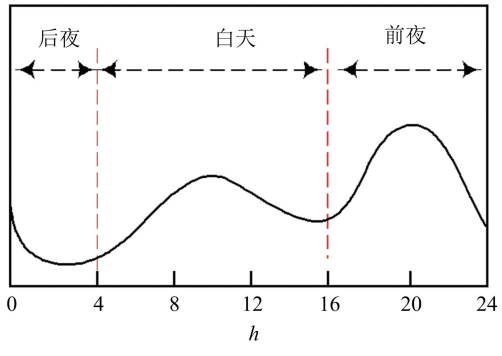

图 2　居民日用电量示意图

Fig. 2　Residential daily electricity consumption

由上述内容可知，数据流上的隐含语义可以辅助不同领域的人员准确地完成工作.

自然语言处理领域中，已有研究工作旨在发现文本中的潜在主题[7-9]，本文将比较成熟的词向量（word embedding）概念[10]扩展到数据流应用中，根据文献[11]的工作，建立数据流的 SP-tree，提取可伸缩模式，并作为特征，建立潜在向量空间，将可伸缩模式向量化，同时也起到了降低特征维度的效果. 通过向量之间的计算，可得到可伸缩模式之间的相似度，然后通过数据流上可伸缩模式的向量组合. 在心电监护领域，可以与心脏疾病对应的向量进行相似度比较，帮助医护人员进行心脏疾病的准确分类工作. 在电网数据应用中，可以用来发现用电异常用户.

本文提出的方法即先提取数据流上的可伸缩模式，并将其向量化表达，发现潜在语义，从而为分类工作提供支持. 相关的工作可以分为以下 3 类：

（Ⅰ）频繁模式挖掘是数据流模式挖掘中最常见的研究方向，主要有两类经典的方法：基于先验原理的方法（apriori based method）[12] 和基于 FP 树（frequent pattern tree）[13]的方法. 频繁模式并一定不能表达复杂的数据流，而且有较高领域价值的潜在语义模式并不一定频繁或者伪周期出现.

(Ⅱ)符号化模式挖掘可以依据某种变化规则,将数值形式的数据流转换成离散的符号序列. 文献[14]提出了一种新型符号化表示方法对时间序列进行符号化近似表示,SAX 是最常用的符号化表示方法. 符号化模式虽然可以用来表达数据流,但是空间代价较高,而且精度牺牲太大.

(Ⅲ)特定模式挖掘,主要用来提取数据流上特定的波段,比如心电监护数据流上经常对 QRS 波群进行提取[15],但是特定模式挖掘的应用较为单一,既不具有普适性,也不能发现数据流上的潜在语义.

1 总体框架

本文基于数据流上的可伸缩模式,提出了 Pattern2vec 的方法,Pattern2vec 的总体框架如图 3 所示. 首先,利用 SP-tree 存储原始数据流,并挖掘出其中的可伸缩模式集合;其次,将可伸缩模式映射到潜在向量空间;最后,根据可伸缩模式的向量表达为分类工作提供辅助.

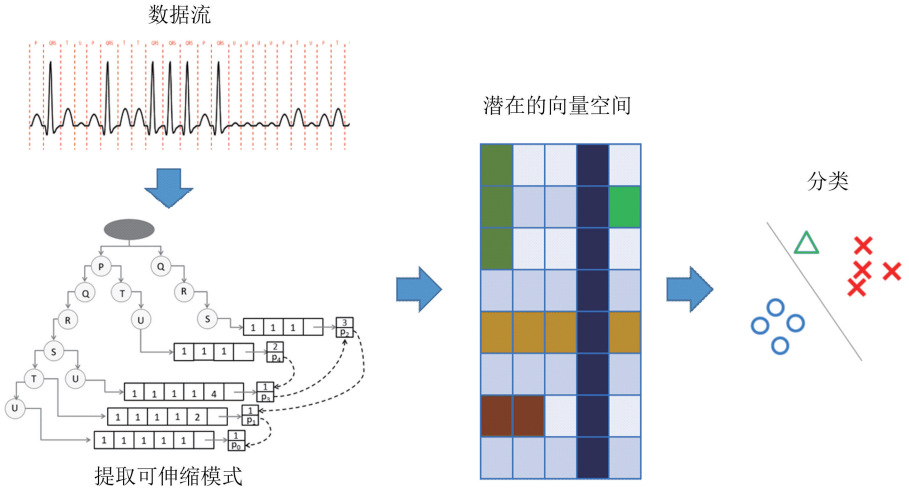

图 3 Pattern2vec 框架图

Fig. 3 Pattern2vec framework

1.1 基本定义

与“词向量”类比,数据流相当于文本中的正常语句,每个可伸缩模式相当于文本中的每个词语. 先从训练样本中得到大量的可伸缩模式,类比词汇个数为 V. 同时建立一个哈希表 h ,key 为可伸缩模式,value 为可伸缩模式 ID. 相关的符号定义如表 1 所示,其中语境表示可伸缩模式前后顺序出现的可伸缩模式集合.

表 1 Pattern2vec 相关符号定义

Tab. 1 Pattern2vec definition of related symbols

符号含义	形式化符号
数据集	E
可伸缩模式	sp
可伸缩模式序列	sps
语境	Context(sp)
语境中的可伸缩模式	i
模型参数	θ
可伸缩模式哈希表	h
可伸缩模式个数	V
可伸缩模式维度	d

1.2 提取可伸缩模式序列

首先基于原始数据流构建 SP-tree,然后遍历 SP-tree,根据 BPT 和 PFNLL[16] 相关信息,按先后顺序读取对应的可伸缩模式. 同时查询 h ,由可伸缩模式查询对应的 key,通过对应的 value 值得到可伸缩模式序列编号,例如某可伸缩模式序列可以表示为 sps_i ,即

$$sps_i = 4 \rightarrow 3 \rightarrow 1 \rightarrow 5 \rightarrow 10 \rightarrow 46 \rightarrow 51 \rightarrow 89.$$

2 可伸缩模式向量化映射

针对不同专业领域的分类任务,本文的期望是提取可伸缩模式序列,交由机器学习算法来自动诊断. 由于机器无法直接理解可伸缩模式序列所对应的疾病到底是什么. 所以本节对这些可伸缩模式进行向量化的映射,不仅可以得到可伸缩模式序列的数学化表示,还可以得到不同可伸缩模式序列之间的关系.

针对上文提取到的可伸缩模式序列 sp_{si} ,假如选定 1 号可伸缩模式,将其映射到对应的向量空间. 本文使用经典的 Skip-gram 模型[17],如图 4 所示.

即根据所选可伸缩模式，预测周围可伸缩模式语境. 同时，针对序列，定义滑动窗口 Ω，在 1 到 5 之间对其进行随机选取. 如果 $\Omega=1$，那 1 号可伸缩模式的语境为 3 号和 5 号可伸缩模式，因此需要

$$\text{Maximize}: p(sp_3 \mid \Phi sp_1), p(sp_5 \mid \Phi sp_1) \quad (1)$$

则本文的目的是最大化目标函数，使得给定可伸缩模式的情况下，可伸缩模式的语境出现的概率最大，并求得此时的参数 θ.

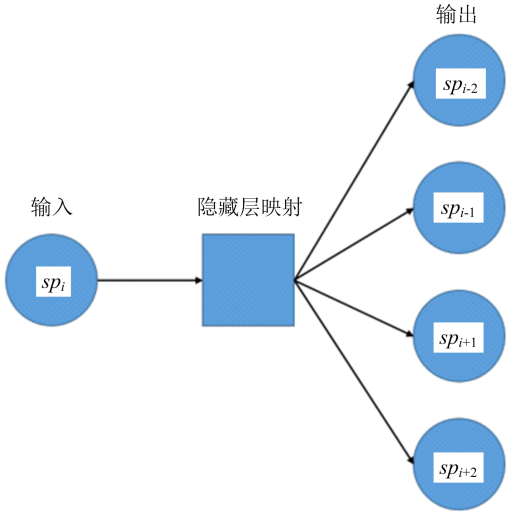

图 4　Skip-gram 模型

Fig. 4　Skip-gram model

条件函数 $p(\text{Context}(sp)\mid sp;\theta)$ 定义为

$$p(\text{Context}(sp)\mid sp;\theta)=\prod_{i\in\text{Context}(sp)} p(i\mid sp;\theta) \quad (2)$$

使用逻辑回归的扩展 Softmax 对 θ 进行形式化处理，使得条件概率转为

$$p(\text{Context}(sp)\mid sp;\theta)=\frac{e^{v_i\cdot v_{sp}}}{\sum_{j\in D} e^{v_j * v_{sp}}} \quad (3)$$

式中，v_i，v_{sp} 分别是 i 和 sp 的列向量，维数为 d；V 是所有心电监护数据中的可伸缩模式构成的集合；参数 θ 是 v_i 和 v_{sp} 中每一个维度的取值，参数的总个数为 $|i| * |sp| * |V|$. 代入公式(3)，两边取对数可以得到

$$\log p(\text{Context}(sp)\mid sp;\theta)=\log e^{v_i\cdot v_{sp}}-\log\sum_{j\in V} e^{v_j * v_{sp}} \quad (4)$$

至此，已经推导出对数似然函数的表达式，这就是本文向量化可伸缩模式的目标函数，接下来就可以利用随机梯度上升法对其进行优化. Skip-gram 模型的隐层到输出层，需要直接计算每一个可伸缩模式的输出概率，即传统的 Softmax 模型，由于可伸缩模式的个数较多，即 V 较大，因此 $|V|$ 中的每一个可伸缩模式都要进行计算，时间复杂度是 $O(|V|)$.

3　层次 Softmax

本文为了降低时间复杂度，使用层次 Softmax，通过统计可伸缩模式的出现频率构建了 Huffman 树，如图 5 所示，具体的算法描述如算法 3.1 所示，时间复杂度由 $O(|V|)$ 降到了 $O(\log_2(|V|))$，因此公式(1)应该改为

$$\text{Maximize}: p(\text{left}\mid \varphi sp_1; C_1);$$
$$\text{Maximize}: p(\text{right}\mid \varphi sp_1; C_2);$$
$$\text{Maximize}: p(\text{left}\mid \varphi sp_1; C_3).$$

算法 3.1　构建 Huffman 树

输入：可伸缩模式频率 spCount 和可伸缩模式种类 n；

输出：Huffman 树的根节点 HuffmanTreeRoot.

1 FOR i←1 到 DO

2 新建由(可伸缩模式的标号 i，可伸缩模式频率 spCount_i)构成的 HuffmanTree 的节点 node_i(左右儿子都为空节点)

3 将节点 node_i 放到优先队列 Q

4 END FOR

5 FOR i←1 到 n−1 DO

6 从优先队列 Q 中挑选出频率最小的两个节点 node_a 和 node_b，并从 Q 中去掉这两个节点

7 新建 node_a 和 node_b 的父节点 node_f(父节点标号 i，node_a 和 node_b 的频率之和 spCount_f)

8 节点 node_f 的左右儿子分别指向 node_a 和 node_b

9 把节点 node_f 放入优先队列 Q 中

10 END FOR

11 取出 Q 中最后的一个节点 HuffmanTreeRoot

12 返回 HuffmanTreeRoot

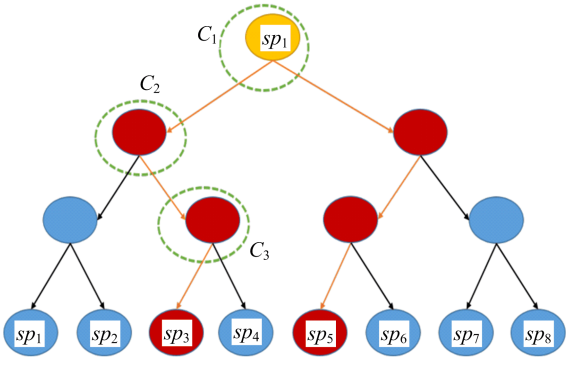

图 5　层次 Softmax 示意图

Fig. 5　Hierarchical Softmax sketch

在进行参数估计的时候,需要先随机初始化可伸缩模式的向量,然后对每一个 C_i 计算代价函数,最后利用随机梯度下降迭代更新分类器的权值以及各个可伸缩模式的向量.

根据层次 Softmax 的思想,公式(2)可以写为

$$p(\text{Context}(sp)\mid sp;\theta)=\prod_{i\in\text{Context}(sp)}p(i\mid sp;\theta)=\prod_{i\in\text{Context}(sp)}\prod_{j=2}^{l^i}p(d_j^i\mid v(sp),\theta_{j-i}^i)\tag{5}$$

式(5)表明,对于任意可伸缩模式 sp,Huffman 树中必存在一条从根节点到词 i 对应结点的唯一路径并存在 l^i-1 个分支,将每个分支看作一次二分类,每一次分类就产生一个概率,将这些概率乘起来就得到了 $p(i\mid sp;\theta)$.其中,

$$p(d_j^i\mid v(sp),\theta_{j-1}^i)=\sigma(v(sp)^T\theta_{j-1}^i),d_j^i=1;$$

$$p(d_j^i\mid v(sp),\theta_{j-1}^i)=1-\sigma(v(sp)^T\theta_{j-1}^i),d_j^i=0.$$

写成整体表达式,即

$$p(d_j^i\mid v(sp),\theta_{j-i}^i)=[\sigma(v(^sp)^T)]^{1-d_j^i}*[1-\sigma(v(^sp)^T)]^{d_j^i}.$$

将以上推导带入公式(5),可以得到对数似然函数的具体表达式.

记上式为

$$\mathcal{L}=\sum_{sp\in E}\sum_{i\in\text{Context}(sp)}\sum_{j=2}^{l^i}\mathcal{L}(sp,i,j)\tag{6}$$

接下来,使用随机梯度上升法进行最优化问题的求解,分别需要考虑关于 θ_{j-1}^i 和 $v(sp)$ 的梯度,推导过程如下:

$$\begin{aligned}\frac{\partial\mathcal{L}(sp,i,j)}{\partial\theta_{j-1}^i}&=\frac{\partial}{\partial\theta_{j-1}^i}\{(1-d_j^i)*\log[\sigma(v(sp)^T\theta_{j-1}^i)]+d_j^i*\log[1-\sigma(v(sp)^T\theta_{j-1}^i)]\}=\\&(1-d_j^i)[1-\sigma(v(^sp)T\theta_{j-1}^i)]v(sp)-d_j^i[\sigma(v(^sp)T\theta_{j-1}^i)]v(sp)=\\&\{(1-d_j^i)[1-\sigma(v(sp)^T\theta_{j-1}^i)]-d_j^i[\sigma(v(sp)^T\theta_{j-1}^i)]\}v(sp)=\\&[1-d_j^i-\sigma(v(^sp)T\theta_{j-1}^i)]v(sp)\end{aligned}\tag{7}$$

$$\begin{aligned}\frac{\partial\mathcal{L}(sp,i,j)}{\partial v(sp)}&=\frac{\partial}{\partial v(sp)}\{(1-d_j^i)*\log[\sigma(v(sp)^T\theta_{j-1}^i)]+d_j^i*\log[1-\sigma(v(sp)^T\theta_{j-1}^i)]\}=\\&(1-d_j^i)[1-\sigma(v(^sp)T\theta_{j-1}^i)]\theta_{j-1}^i-d_j^i[\sigma(v(^sp)T\theta_{j-1}^i)]\theta_{j-1}^i=\\&\{(1-d_j^i)[1-\sigma(v(sp)^T\theta_{j-1}^i)]-d_j^i[\sigma(v(sp)^T\theta_{j-1}^i)]\}\theta_{j-1}^i=\\&[1-d_j^i-\sigma(v(^sp)T\theta_{j-1}^i)]\theta_{j-1}^i\end{aligned}\tag{8}$$

因此,更新公式可以分别写为

$$\theta_{j-1}^i=\theta_{j-1}^i+\eta[1-d_j^i-\sigma(v(sp)^T\theta_{j-1}^i)]v(sp)\tag{9}$$

$$v(sp)=v(sp)+\eta\sum_{sp\in E}\sum_{j=2}^{l^i}[1-d_j^i-\sigma(v(sp)^T\theta_{j-1}^i)]\theta_{j-1}^i\tag{10}$$

整个训练过程如算法 3.2 所示.

算法 3.2 练 Skip-gram

输入:可伸缩模式序列 sps、可伸缩模式列向量维度 d、Huffman 树 HuffmanTree、最大窗口 maxWindow,最大训练轮数 maxEpoch,可伸缩模式词典 V;

输出:每个可伸缩模式的隐向量 v.

```
随机初始化输入层到隐藏层的|V|×d 矩阵 σ
随机初始化隐藏层到输出层的|V|×d 矩阵 θ
FOR epoch←1 到 maxEpoch DO
  FOR SPS 中每一个可伸缩模式 sp_i DO
    window←1 到 maxWindow 的随机数
    e←零向量
    FOR j←1 到 window DO
      FOR Huffman 树的根节点到叶子节点 sp_{i+j} 的路径上的非叶子节点 node_k DO
        IF node_k 为其父节点的左儿
          L←0
        ELSE
          L←1
        END IF
        q←σ(v(sp_i)^T · θ^k)
        g←η(L-q)
        e←e+gθ^k
        θ^k←θ^k+g · v(sp_i)
      END FOR
    END FOR
    v(sp_i)←v(sp_i)+e
  END FOR
END FOR
返回输入层到隐藏层的矩阵 v
```

4 判别分类

上文对数据流上提取的可伸缩模式序列进行了向量化映射,并得到了可伸缩模式序列对应的向量表示,本节将会基于这些向量进行分类,比如基于心

电监护数据对病人所患心脏疾病的种类进行分类.

首先对距离进行定义. 设 sp_x,sp_y 是来自协方差矩阵 $\boldsymbol{\Sigma}$(均值向量为 μ)的两个向量化可伸缩模式,则 sp_x,sp_y 之间距离为马氏平方距离,表示为

$$d^2(sp_x,sp_y)=(sp_x,sp_y)^T\sum{}^{-1}(sp_x-sp_y) \tag{11}$$

定义 sp 与总体 E 的距离

$$d^2(sp,E)=(sp-\mu)^T\sum{}^{-1}(sp-\mu) \tag{12}$$

假设有 N 种心脏疾病,第 i 种疾病对应的可伸缩模式向量集合用 E_i 表示,均值向量和协方差矩阵分别为 μ_i 和 E_i ,通过计算病人心电监护数据向量 v 到各总体的距离,比较这些距离,判定 v 属于与其马氏距离最小的总体.

对于任意两种疾病 E_i ,E_j ,考虑 v 到 E_i 和 E_j 的距离的差

$$d^2(v,E_i)-d^2(v,E_j)=-2[W_j(v)-W_i(v)] \tag{13}$$

其中,

$$W_i(v)=a_i^Tv+b_ia_i=\sum{}^{-1}\mu_ib_i=-\frac{1}{2}\mu_i^T\sum{}^{-1}\mu_i;$$

$$W_j(v)=a_j^Tv+b_ja_j=\sum{}^{-1}\mu_jb_j=-\frac{1}{2}\mu_j^T\sum{}^{-1}\mu_j;$$

因此

$$d^2(v,E_i)\geqslant d^2(v,E_j)\Leftrightarrow W_i(v)\leqslant W_j(v).$$

这样,判定病人心电监护数据向量 v 属于第 m 类,满足

$$W_m(v)=\max_{1\leqslant n\leqslant N}W_n(v).$$

5　实验结果与分析

5.1　实验数据集

本实验的数据集来源于医疗数据和电力数据. 基于这两类数据集进行实验,并比较它们的分类效果.

医疗数据来自两个数据库,一个为 MIT-BIH 数据库,从中选取了 3 种心脏疾病的数据,分别为:室性快速型心律失常(cuvtd)、室性心律失常(mvad)和心律不齐(madd). 该数据库由 48 个经过注解的记录组成,每个记录时间约为 30 分钟,数据的采集频率为 360 Hz;另一个为 MIMIC 数据库,从中抽取了四类心脏病患者的心电监护数据,分别为早搏(PB)、纤颤(FI)、冠状动脉功能不全(CI)和心肌炎(MC),该数据集采集频率为 125 Hz.

电力数据来自于某电网公司近 15 万名用户在 2014 年至 2016 年每天的用电量数据,数据包含日期、当天电能表示值,前一天电能表示值和用电量数值,这些用户包含窃电用户和非窃电用户.

5.2　隐含语义发现

本实验基于 Pattern2vec 的方法,发现了可伸缩模式间和心电监护数据整体间的隐含语义.

基于 MIT-BIH 数据集,利用 Pattern2vec 方法得到了对应的可伸缩模式向量表达,由于可伸缩模式向量维度为 50,因此为了更好地可视化体验,本文采用文献[18]提出的 t-SNE 方法,对高维向量进行二维可视化展示,图 6 展示了部分实验结果. 由图 6 可以发现,$PQRST^3$ 和 $PQR(STT)^3$ 空间距离较近,说明它们相似度较高. $(PQRS)^3$ 和 $(PQRS)^4$ 空间距离较近,RST、QRSU 和 RSTT 三者距离较近. 验证了前文提到的可伸缩模式之间,虽然表现不同,但是存在隐含的相似语义.

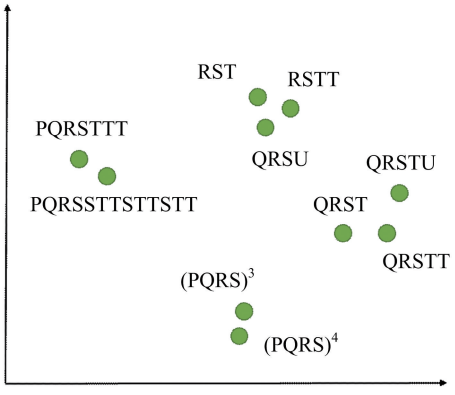

图 6　可伸缩模式向量空间示意图

Fig. 6　Vector space of scalable pattern

同时也发现 QRST、QRSTU 和 QRSTT 相似度高,本文猜测可能是心电监控数据采集的时候,有噪声带入,造成了可伸缩模式的不同表达.

本实验基于 MIT-BIH 和 MIMIC 数据集,分别得到了心电监护数据上可伸缩模式的向量,把每个病患的可伸缩模式向量进行线性相加,便可得到该病患心电监护数据的整体向量表达. 图 7 是 3 种疾病的心电监护数据在二维空间映射实验结果,图 8 是 4 种疾病的心电监护数据在二维空间映射实验结果. 可以发现,在 MIT-BIT 数据集上,cutvd 和 madd 数据集聚集效果较为明显,mvad 分布相对比较分散. MIMIC 数据集上的 4 种心脏疾

病对应的心电监护数据具有更为明显的聚集效应. 从实验结果可以证明心电监护数据整体具有隐含语义.

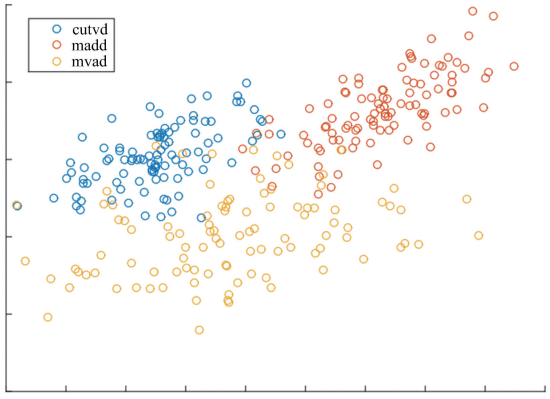

图 7 MIT-BIT 数据集上心电监控数据向量空间示意图

Fig. 7 Vector space of ECG monitoring data on MIT-BIT dataset

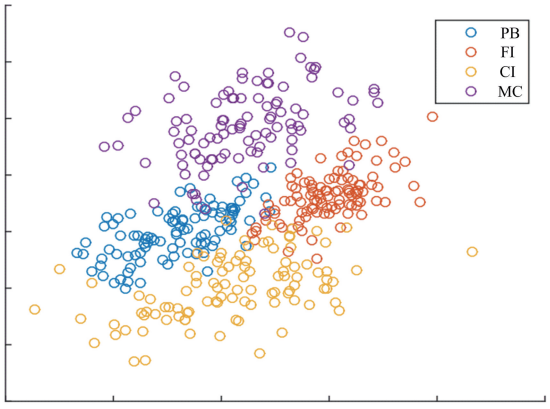

图 8 MIMIC 数据集上心电监控数据向量空间示意图

Fig. 8 Vector space of ECG monitoring data on MIMIC dataset

5.3 医疗数据流分类效果

基于心电监护数据,利用 Pattern2vec 的方法对心脏疾病进行分类实验. 同时也选择 KNN+DTW[19],CNN[20] 和 SVM 为对比方法,其中 SVM 为经典的多分类器,本实验将可伸缩模式当作 SVM 方法的特征,进行相关实验.

本实验利用 Micro-F_1 和 Macro-F_1 衡量各方法的分类效果,实验结果如表 2 和表 3 所示. 随着训练集比例的增大,F_1 值均呈增长趋势. 可以很明显地发现,相比于对比方法,Pattern2vec 方法在两组数据集上均有更好的分类效果. 另外,由于本实验的 SVM 方法将可伸缩模式作为特征,因此与 KNN+DTW 和 CNN 的实验结果比较而言,本实验中的 SVM 方法分类效果更好.

表 2 MIT-BIT 数据集上的分类 F_1 值

Tab. 2 Classification F_1 on MIT-BIT dataset

	%训练集	40%	50%	60%	70%
Micro-F_1 (%)	Pattern2vec	37.52	39.26	39.69	42.46
	KNN+DTW	23.27	23.26	28.11	30.42
	CNN	29.22	33.73	36.13	38.27
	SVM	34.22	35.29	39.11	40.28
Macro-F_1 (%)	Pattern2vec	31.22	32.23	33.13	35.56
	KNN+DTW	17.26	19.02	17.66	22.21
	CNN	22.20	22.46	23.58	23.66
	SVM	21.11	26.45	28.23	30.22

表 3 MIMIC 数据集上的分类 F_1 值

Tab. 3 Classification F_1 on MIMIC dataset

	%训练集	40%	50%	60%	70%
Micro-F_1 (%)	Pattern2vec	35.24	38.15	41.25	43.21
	KNN+DTW	20.12	26.33	24.56	30.22
	CNN	26.54	30.85	37.45	39.61
	SVM	32.22	34.25	39.11	40.25
Macro-F_1 (%)	Pattern2vec	30.11	32.56	33.21	34.56
	KNN+DTW	16.78	15049	16.34	22.59
	CNN	20.59	22.65	24.85	26.59
	SVM	26.58	28.56	30.13	32.36

5.4 电力数据流分类效果

对是否为窃电用户进行预测,可以看作一个二分类的问题,窃电用户为正样本,非窃电用户为负样本,本节将对 Pattern2vec、SAX[21]、ADAP[22] 和 SVM-kNN[23] 的方法进行性能评估. 引入混淆矩阵[24]的概念,如表 4 所示,混淆矩阵中的实例可以分为以下 4 类:

表 4 混淆矩阵

Tab. 4 Confusion matrix

	P'预测为真	N 预测为假
P 真实为真	TP	FN
N'真实为假	FP	TN

(ⅰ)真正(true positive, TP):被预测为正的正样本;

(ⅱ)假正(false positive, FP):被预测为正的负样本;

(ⅲ)假负(false negative, FN):被预测为负的

正样本；

(ⅳ)真负(true negative，TN)：被预测为负的负样本.

实验首先对准确率进行分析，假设预测方法能把更多的正样本预测为正，更多的负样本预测为负，则该方法的准确率较高，因此准确率的计算方法为

$$\text{accuracy}=(\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{FP}+\text{TN}) \tag{14}$$

由于非窃电用户的用电习惯比较类似，因此本文的 Pattern2vec 方法对大概率事件样本采用了不同的采样比例(η). 基于已有经验，设 $\eta=0.01$，0.02，0.1，进行了多次试验. ADAP 方法也选择了 3 个效果最好的参数($d=30,50,70$)进行准确率对比.

为了更好地分析不同方法的性能，本文针对分类结果进行了 ROC 和 AUC 分析，ROC 图的横坐标为 false positive rate(FPR)，纵坐标为 true positive rate(TPR)，计算公式为

$$\text{FPR}=\text{FP}/(\text{FP}+\text{TN}) \tag{15}$$

$$\text{TPR}=\text{TP}/(\text{TP}+\text{FN}) \tag{16}$$

本实验根据上述每种方法的分类结果绘制了 ROC 曲线，如图 9 所示，显示了各方法在本实验电力数据集上的分类表现.

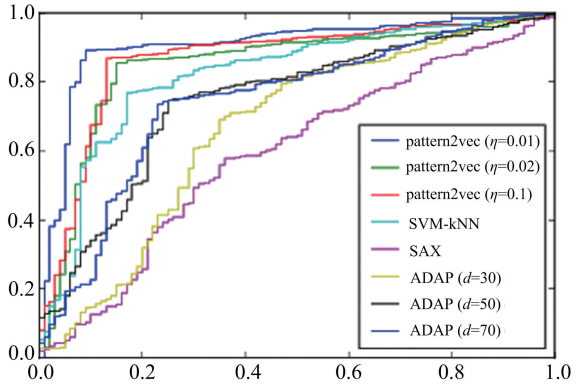

图 9　电力数据上不同方法的 ROC 图

Fig. 9　ROC of different methods on electrical data

本实验计算了各方法对应 AUC 值(即 ROC 曲线下方的面积)，结果如表 5 所示. 由表 5 可以看出，Pattern2vec 方法的 AUC 值大于其他对比方法，且所有 AUC 值都在 0.9 左右，这意味着 Pattern2vec 具有非常好的预测效果. 同样，利用可伸缩模式作特征的 SVM-kNN 也具有较高的 AUC 值.

表 5　不同方法的 AUC 值比较

Tab. 5　AUC comparison in different methods

方法	AUC 值
Pattern2vec($\eta=0.01$)	0.885
Pattern2vec($\eta=0.02$)	0.902
Pattern2vec($\eta=0.1$)	0.916
SVM-kNN	0.823
SAX	0.614
ADAP(d=30)	0.695
ADAP(d=50)	0.715
ADAP(d=70)	0.781

6　结论

本文分析了复杂数据流上的语义隐含特点，提出了 Pattern2vec 的方法. 首先，利用 SP-tree 来存储原始数据流，并挖掘出其中的可伸缩模式集合；其次，将可伸缩模式映射到潜在向量空间；最后，根据可伸缩模式的向量表达为分类工作提供辅助. 本文选择了医疗和电力领域，分别基于 3 类心脏病(心肌梗死、心律不齐和冠心病)、4 类心脏病(早搏、纤颤、冠状动脉功能不全和心肌炎)的患者心电监护数据以及某电网居民用电数据进行了不同的分类实验. 实验结果证明，Pattern2vec 的确能发现数据流上的隐含语义事件，同时，与对比方法的实验结果相比，Pattern2vec 具有更好的分类效果.

参考文献(References)

[1] EOM J H, KIM S C, ZHANG B T. AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction[J]. Expert Systems with Applications, 2008, 34(4): 2465-2479.

[2] JONES A. Assessment Problems-Clinical Electrocardiography: A Simplified Approach (Seventh Edition)-Self [M]// Australian Journal of Anthropology, 2015, (3):365-380.

[3] BANERJEE S, MITRA M. Application of cross wavelet transform for ECG pattern analysis and classification [J]. IEEE Transactions on Instrumentation & Measurement, 2014, 63(2): 326-333.

[4] KIM K K, LIM Y G, KIM J S, et al. Effect of missing RR-interval data on heart rate variability

analysis in the time domain [J]. Physiological Measurement, 2007, 28(12):1485-1494.
[5] CHINNASAMI S, RATHORE C, DUNCAN J S. Sinus node dysfunction: An adverse effect of lacosamide[J]. Epilepsia, 2013, 54(6):90-93.
[6] 李菲菲, 李红燕, 曲强, 等. SPQ:数据流上面向可伸缩模式的查询方法[J]. 计算机学报, 2010, 33(8): 1481-1491.
[7] XU W, RUDNICKY A. Can artificial neural networks learn language models? [C]// Proceedings of 6th International Conference on Spoken Language Processing. Beijing: IEEE, 2000:202-205.
[8] BLEI D M, NG A Y, JORDAN M I. Latent Dirichletal location[J]. Journal of Machine Learning Research Archive, 2003, 3: 993-1022.
[9] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1-2):177-196.
[10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.
[11] QIU Z, LI FF, HONG S D, et al. A novel method for mining semantics from patterns over ECG data[C]// Workshops at the 30th AAAI Conference on Artificial Intelligence. Hyatt, USA: ACM, 2016: 446-452.
[12] AGRAWAL R, SRIKANT R. Fast Algorithms for Mining Association Rules[M]// Readings in database systems 3ed, Morgan Kaufmann Publishers Inc., 1998.
[13] WANG L, CHEUNG D W, CHENG R, et al. Efficient mining of frequent item sets on large uncertain databases [J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24 (12): 2170-2183.
[14] LIN J, KEOGH E, LONARDI S, et al. A symbolic representation of time series, with implications for streaming algorithms[C]// ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego: ACM, 2003:2-11.
[15] MITRA S, MITRA M, CHAUDHURI B B. Pattern defined heuristic rules and directional histogram based online ECG parameter extraction[J]. Measurement, 2009, 42(1):150-156.
[16] 巨程, 李红燕, 李菲菲, 等. OCPM:一种数据流上复合模式的在线匹配方法[J]. 计算机研究与发展, 2011, 48(3):304-311.
[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J/E]. OALib Journal, Computer Science, 2013: arXiv:1301.3781v3.
[18] LAURENS V D M, HINTON G, HINTON V D M G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9 (2605): 2579-2605.
[19] RAKTHANMANON T, CAMPANA B, MUEEN A, et al. Searching and mining trillions of time series subsequences under dynamic time warping [C]// International Conference on Knowledge Discovery and Data Mining. Beijing: ACM, 2012:262-270.
[20] KIRANYAZ S, INCE T, HAMILA R, et al. Convolutional neural networks for patient-specific ECG classification [C]//37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Milan Italy: IEEE, 2015:2608-2611.
[21] LIN J, KEOGH E, WEI L, et al. Experiencing SAX: A novel symbolic representation of time series[J]. Data Mining & Knowledge Discovery, 2007, 15(2): 107-144.
[22] CHIA C C, SYED Z. Scalable noise mining in long-term electrocardiographic time-series to predict death following heart attacks[C]// Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2014: 125-134.
[23] NAYAK R K, MISHRA D, RATH A K. A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices [J]. Applied Soft Computing, 2015, 35:670-680.
[24] TOWNSEND J T. Theoretical analysis of an alphabetic confusion matrix[J]. Perception & Psychophysics, 1971, 9(1):40-50.
[25] TSENG P. Convergence of a block coordinate descent method for nondifferentiableminimization[J]. Journal of Optimization Theory & Applications, 2001, 109 (3):475-494.
[26] HARPER F M, KONSTAN J A. The MovieLens datasets: History and context[J]. ACM Transactions on Interactive Intelligent Systems, 2015,5(4): 1-19.