

基于子序列相似性的时间序列语义挖掘算法

陆 怡¹, 王 鹏², 汪 卫²

(1. 复旦大学 软件学院, 上海 201203; 2. 复旦大学 计算机科学技术学院, 上海 201203)

摘 要: 时间序列是对某个事物或系统进行连续同间隔测量得到的数值序列, 挖掘时间序列中潜在的语义信息对于发现系统运行规律或识别系统突发异常至关重要, 然而目前多数时间序列语义挖掘算法对于时间序列数据特征有一定的约束条件, 难以处理海量且特征各异的时间序列数据。针对该问题, 提出一种基于子序列相似性的时间序列语义挖掘算法。通过计算子序列的相似性, 将时间序列分割成片段序列进行两级聚类, 识别出时间序列中潜在的物理状态。引入基于概率的迭代模式, 根据候选分段情况动态调整子序列被选为参考子序列的概率, 保证参考子序列涵盖全部物理状态。实验结果表明, 该算法在PAMAP、Barbet等5个真实数据集上的识别准确率均超过90%, 相比于FLUSS、pHMM、AutoPlait算法具有更高的识别准确率与运行效率以及更强的通用性。

关键词: 时间序列; 语义挖掘; 相似性度量; 聚类; k最近邻

开放科学(资源服务)标志码(OSID):



中文引用格式: 陆怡, 王鹏, 汪卫. 基于子序列相似性的时间序列语义挖掘算法[J]. 计算机工程, 2022, 48(10): 88-94.

英文引用格式: LU Y, WANG P, WANG W. Time-series semantic mining algorithm based on sub-series similarity[J]. Computer Engineering, 2022, 48(10): 88-94.

Time-Series Semantic Mining Algorithm Based on Sub-Series Similarity

LU Yi¹, WANG Peng², WANG Wei²

(1. School of Software, Fudan University, Shanghai 201203, China;

2. School of Computer Science, Fudan University, Shanghai 201203, China)

[Abstract] Time-series is a sequence of values obtained by continuously measuring an object or system at the same interval. By obtaining potential semantic information in the time-series, the regularities or anomalies of a system can be identified, which can provide guidance for practice and analysis. However, most current time-series semantic mining algorithms are constrained by some of the characteristics of time-series data, and addressing a significant amount of time-series data with different characteristics is difficult. Hence, a time-series semantic mining algorithm based on sub-series similarity is proposed herein. First, by calculating the similarity of sub-series, the algorithm partitions the time-series into segment sequences for two-level clustering and identifies the underlying physical states in the time-series. Second, the algorithm introduces an iterative mode based on probability, dynamically adjusts the probability of a sub-series selected as a reference sub-series based on the candidate segmentation, and ensures that the reference sub-series includes all physical states. Experimental results show that the recognition accuracy of the algorithm on five real data sets such as PAMAP and Barbet exceeds 90%. Compared with FLUSS, pHMM, and AutoPlait algorithms, the proposed algorithm demonstrates higher recognition accuracy, operating efficiency, and versatility.

[Key words] time-series; semantic mining; similarity measurement; clustering; k Nearest Neighbor(kNN)

DOI: 10. 19678/j. issn. 1000-3428. 0062832

0 概述

时间序列数据是在一段时间内以固定的时间间隔采集的数据点序列, 用于描述现象随时间变化的情况, 已成为日常生活中重要的信息记录形式。随着大数据时代的到来以及大规模计算能力的提升,

时间序列数据分析与挖掘已成为研究热点, 广泛应用于医疗、交通、金融等领域。在现实场景中, 时序数据虽然存在不同的演化规律, 例如记录服务器CPU使用情况的监控数据、反映患者健康状况的心电数据以及表征市场行情的商品销售数据, 但这些领域各异的数据的内核通常是一致的, 会根据观测

基金项目: 国家重点研发计划(2020YFB1710001)。

作者简介: 陆 怡(1997—), 女, 硕士研究生, 主研方向为时间序列数据分析与挖掘; 王 鹏、汪 卫, 教授、博士生导师。

收稿日期: 2021-09-28 修回日期: 2021-11-26 E-mail: luyi19@fudan.edu.cn

对象或系统的实际物理状态,呈现出不同的波动形式。因此,挖掘时间序列中潜在的语义信息,实际上是识别被监测系统的物理状态,通过将时间序列转换为状态序列,实现数据压缩或者异常检测。

时间序列具有数据量大、复杂度高、干扰信息多等特性,一般采用特征提取手段对时间序列进行加工处理。针对时间序列的语义挖掘,主要分为基于领域特征、基于全局特征、基于子序列相似性三类。基于领域特征的时间序列语义挖掘算法除了时间序列数据之外,还引入了领域知识或者带标签的数据。文献[1]针对金融时间序列的自相关特征,提出一种基于ARMA模型^[2]的时间序列分割算法。文献[3]根据心电图时间序列的特点,提出一种基于残差平衡及边界约束的分段线性回归方法以得到时间序列的分段表示。文献[4]通过从训练集中构建特征的高斯概率分布模型,实现有监督的状态挖掘。基于全局特征的时间序列语义挖掘算法使用全局特征对时间序列进行概括。全局特征又可细分为斜率和均值两种类型。一种以pHMM^[5]为代表,从斜率的角度将时间序列简化成线段序列,并利用隐马尔科夫模型(Hidden Markov Model, HMM)^[6]指导线段的聚类,最终获得状态信息。另一种以AutoPlait^[7]和StreamScope^[8]为代表,从均值的角度描述不同的状态,得到多层的隐马尔科夫模型,并通过最小描述长度(Minimum Description Length, MDL)^[9]对可选的模型进行评估,从而得到时间序列的最优表示。基于均值的算法还可进一步应用于多维或高阶时间序列,例如DBSE^[10]以及在AutoPlait基础上提出的CUBEMARKER^[11]。在基于子序列相似性的时间序列语义挖掘算法中,具有代表性的子序列会频繁出现。实际上,大量对于时间序列的分析是从子序列入手的,例如文献[12-13]实现的基序列挖掘,文献[14]提出的基于子序列的时间序列分类方法。但是,基于子序列相似性的语义挖掘算法仍在初步研究阶段。FLUSS^[15-16]和ESPRESSO^[17]作为典型代表,仅能识别出状态的转变点。

目前,多数时间序列语义挖掘算法对于时序数据特征有一定的约束条件,因此难以处理海量且形式各异的时序数据。本文提出一种基于子序列相似性的时间序列语义挖掘算法SEDIS,通过计算子序列的相似性,将时间序列分割成片段序列进行两级聚类,识别出时间序列中潜在的物理状态,使其具备在不同应用场景下的通用处理能力。

1 问题描述

定义1(时间序列) 时间序列是对某个事物或系统进行连续同间隔的测量得到的数值序列,可表示为 $T=(t_1, t_2, \dots, t_L)$,其中, L 为时间序列长度。

定义2(子序列) 子序列是由时间序列中一段连续的值组成的数值序列。 $T_{l:l}=(t_l, t_{l+1}, \dots, t_{l+l-1})$ 表

示从第 l 个时间点开始,长度为 l 的子序列。

时间序列作为事物或系统某个剖面的观测结果,按照时间序列反映的物理意义将其片段组成若干分段,并根据分段的内在联系形成状态序列,称为分段表示与状态表示。

定义3(分段表示) 时间序列 T 的分段表示可定义为 $SEG(T)=\{s_1, s_2, \dots, s_n\}$,其中 $s_i(1 \leq i \leq n)$ 是 T 的一个子序列。

定义4(状态表示) 时间序列 T 的状态表示在分段表示的基础上,引入了分段的物理状态,可定义为 $STATE(T)=\{<s_1, c_1>, <s_2, c_2>, \dots, <s_n, c_n>\}$,其中 c_i 表示分段 s_i 的状态。假设 m 表示物理状态个数,则 c_i 是一个介于1和 m 之间的正整数,且 $1 \leq m \leq n$ 。

根据上文描述,SEDIS的目标任务是寻找一个符合用户期望的状态表示 $STATE(T)$ 以描述或压缩原时间序列 T ,其中包含的物理状态个数 m 由用户定义。实际上,如果将时间序列中的每个时间点视为一个对象,该任务也可理解为聚类任务,目标是将反映同一状态的观测值聚集在同一类簇中。

2 时间序列语义挖掘算法

SEDIS基于以下基本假设:在同一状态下,由于具有代表性的子序列会频繁出现,因此属于同一状态的分段之间必然存在多个相似子序列。SEDIS分为两个阶段:1)基于子序列相似性结合基于密度的聚类方法构建组成分段的片段;2)采用贪心策略对片段进行预聚类,再通过k-means聚类识别出片段对应的物理状态。

2.1 片段分割

为避免两两计算子序列相似性带来的巨大计算开销,SEDIS采用基于概率的抽样方法,每轮选取一个子序列作为参考子序列。如果两个子序列相似,则它们与参考子序列的距离相近。

2.1.1 子序列相似性的优化计算

子序列相似性通过标准化欧式距离进行度量。在欧式距离的基础上,预先对每个子序列进行标准化操作,使得子序列各个时间点对应的数值的均值等于0、标准差等于1,从而消除了幅度缩放、基线漂移等对波形相似性的影响。具体而言,两个长度等于 l 的子序列 $X=(x_1, x_2, \dots, x_l)$ 和 $Y=(y_1, y_2, \dots, y_l)$ 的标准化欧式距离可通过式(1)计算:

$$\text{Dist}(X, Y) = \sqrt{\sum_{i=1}^l \left(\frac{x_i - \mu_X}{\sigma_X} - \frac{y_i - \mu_Y}{\sigma_Y} \right)^2} \quad (1)$$

其中: μ 表示均值; σ 表示标准差。

假设时间序列的长度是 L ,参考子序列的长度是 l ,需要计算参考子序列与其余 $L-l+1$ 个长度同样等于 l 的子序列的标准化欧式距离。由于时间序列的数据量通常较大,因此SEDIS采用文献[18]提出的基于快速傅里叶变换^[19]的算法,加速子序列的相似

性计算。首先对时间序列和参考子序列使用快速傅里叶变换等操作得到两者的滑动点积,然后以滑窗的形式增量地记录子序列的平均值和标准差,从而基于 $O(L\log L)$ 的时间复杂度求出 $L-l+1$ 个子序列对的标准化欧式距离。

由于相似是一个主观的评判标准,难以通过量化后的数值完成二分类,因此为得到多个相似子序列以支持分段,通过最大堆筛选出与参考子序列距离最近的 k 个子序列完成初过滤。

2.1.2 基于密度的优化聚类

由于参数 k 的取值将极大影响 k 个子序列的相似情况,但时间序列的平滑特性以及全自动的应用要求,难以通过类似手肘法的方案提供一个 k 的建议值,因此SEDIS采用基于密度的聚类方法^[20]对 k 个子序列进行再过滤,其核心思想为属于同一分段的子序列在时间轴上会紧密地聚集在一起。基于密度的聚类方法能够自动地将邻近的子序列分到一个类中,进而识别出潜在噪声。具体地,聚类对象是 k 个子序列,聚类相似性标准取决于子序列的起始时间点。子序列对 $T_{i,l}$ 和 $T_{j,l}$ 的距离可简单表示为 $|i-j|$,即两个子序列在时间轴上越接近,越有可能聚成一类。

针对该距离度量,本文提出一种基于密度的优化聚类方法。在确定每个子序列 $T_{i,l}$ 是否为核心点时,需要检索其 E_{Eps} 邻域内存在的子序列数量。由于相似性仅考虑起始时间点,因此问题可简化成找到满足 $\{T_{j,l} | i - E_{\text{Eps}} \leq j \leq i + E_{\text{Eps}}\}$ 的子序列集合。假设子序列已经按照起始时间点升序排序,则可以通过两次二分搜索快速定位第一个满足 $j \geq i - E_{\text{Eps}}$ 的子序列和最后一个满足 $j \leq i + E_{\text{Eps}}$ 的子序列。两者之间包含的子序列数量即 $T_{i,l}$ 的 E_{Eps} 邻域内存在的子序列数量。通过二分搜索,该步骤的时间复杂度可降至对数级。

经过基于密度的聚类方法过滤掉一部分的噪声子序列后,剩余的子序列间可能存在重叠。为便于后续处理,在每一个形成的类簇中,找出其中起始时间点最小的子序列 $T_{i,l}$ 和起始时间点最大的子序列 $T_{j,l}$,以此构建一个新的子序列 $T_{i,j+l-i}$,该子序列即为一个候选分段。

2.1.3 基于概率的迭代模式

由单个参考子序列得到的候选分段仅能表征一个状态,因此需要引入多个不同的参考子序列使其涵盖所有的物理状态。本文提出一种基于概率的迭代模式,根据候选分段的情况动态地调整子序列被选为参考子序列的概率,设计思想为对于没有被包含在候选分段中的子序列给予更多的机会。具体地,对于所有的子序列,赋予相同的初始概率,以模拟均匀分布。在每一轮完成后,将包含在候选分段中的子序列的概率减半,从而直接影响下一轮参考子序列的选取情况。

至此,SEDIS产生若干候选分段,一组候选分段对应一个参考子序列,多组候选分段间可能存在重

叠。候选分段的起始点和终止点表示候选的状态转变点。根据这些候选点对所有候选分段进行分割,产生的子序列称为片段。一个分段至少产生一个片段。值得注意的是,片段不一定足以覆盖整个时间序列,对于漏检片段的处理,将在下文进行具体介绍。

2.2 状态识别

状态识别是根据片段反映的物理状态对其进行聚类,同一类簇的片段对应同一状态。因此,状态识别的核心为片段在物理状态层面上的相似性定义。

2.2.1 基于贪心策略的预聚类

在聚类前,时间序列通常是平滑演进的。排除监测设备的突发故障,时间序列在状态转变点附近的波动通常较小,因此会导致附近的子序列被分割成多个片段。为提高效率,本文对所有生成的片段采用贪心策略进行预聚类,使用栈模拟贪心策略的执行,具体步骤如下:

- 1) 将所有片段按照起始点的降序依次推入栈中。
- 2) 从栈中推出栈顶片段 $T_{i,p}$,找出所有包含该片段的候选分段 $C = \{T_{j,q} | j \leq i < i+p \leq j+q\}$ 。
- 3) 计算集合 C 中所有候选分段的终止点的平均值 $A = \frac{\sum_{T_{j,q} \in C} j+q-1}{|C|}$,其中 $|C|$ 表示集合 C 的元素个数。
- 4) 从 C 中选择终止点最接近 A 的候选分段。将 $T_{i,p}$ 的起始点 i 和选中的候选分段的终止点 $j+q-1$ 组合形成新的片段 $T_{i,j+q-i}$ 。
- 5) 从栈中推出新的栈顶片段,如果被 $T_{i,j+q-i}$ 包含,则略过;否则,重复步骤2~步骤5直到栈中无剩余元素。

基于上述步骤,在保留片段原有语义信息的同时,通过片段间的合并减少了片段数量,提高了算法运行效率。

2.2.2 基于k-means的再聚类

根据聚类的任务描述:将没有分类标签的数据集分为若干个簇^[21],再结合状态识别目标,假设给定一个在物理状态层面上的相似性度量,通过聚类可将片段组织成多个类簇,每个类簇对应一个状态。

SEDIS的基本假设为属于同一状态的片段之间存在多个相似的子序列,因此基于片段分割的过程,提出一种新的相似性度量函数。为便于描述,给定一个片段 P_i ,辅助向量 O_i 是一个 r 维的向量,其中 r 是参考子序列的总个数。 O_i 的每一维是0或者1,表示 P_i 是否被由该参考子序列生成的候选分段包含,进而定义两个片段 P_i 和 P_j 的相似性,如式(2)所示:

$$\text{SIM}(P_i, P_j) = \frac{\text{sum}(O_i \& O_j)}{\min(\text{sum}(O_i), \text{sum}(O_j))} \quad (2)$$

其中: $\&$ 表示按位与;sum表示求和;min表示求最小值。

式(2)计算的是两个片段在同一候选分段中的共现频率,在一定程度上反映了两个片段包含的相似子序列的数量。该值越大,表明两者属于同一状态的可信度越大。

基于相似性定义,采用k-means聚类^[22]算法得到片段的隐含结构,即片段的物理状态,具体步骤如下:

1)随机选取 m 个片段作为初始的聚类中心,其中 m 是用户指定的物理状态个数。

2)将每个片段分配到相似性最高的聚类中心所在的类簇中。

3)对于每个类簇,选取其中与所有其他片段相似性之和最大的片段作为新的聚类中心。

4)重复步骤2和步骤3,直到聚类中心不再变化。

至此,每个类簇对应一个物理状态。

2.2.3 基于k最近邻的漏检片段分类

需要指出的是,虽然SEDIS在片段分割阶段通过基于概率的迭代模式尽可能地保证参考子序列涵盖全部的物理状态,但是片段的漏检现象仍可能存在,即在时间序列中,有部分子序列并未包含在任何候选分段中。对于这部分子序列,或称为漏检片段,SEDIS采用k最近邻(k Nearest Neighbor, kNN)算法进行分类处理。在实际应用中,将每个漏检片段作为参考子序列,考察在标准化欧式距离度量下与其最相似的 k 个子序列的物理状态,取频次最高的作为分类标签。

2.3 时间复杂度分析

SEDIS主要包括片段分割和状态识别两个阶段。理论上,第一阶段的时间复杂度与时间序列的长度强相关,第二阶段的时间复杂度与片段的数量强相关,后者远小于前者,同时在实际应用中证明,第一阶段的时间开销在总时间开销中占比较大,因此本节将主要围绕该阶段展开分析。

片段分割包括对每个参考子序列执行相似性计算、相似子序列筛选、聚类3个步骤。假设时间序列的长度等于 n ,参考子序列的个数等于 r ,每次只保留 k 个与参考子序列最相似子序列,则该阶段的时间复杂度为 $O(r \times (n \log_a n + n \log_a k + k \log_a k))$ 。但需要指出的是,由执行快速傅里叶变换引入的 $O(n \log_a n)$ 的时间复杂度在实际应用中是近似线性的,文献[18]认为这可能是由于该算法已在工程领域中得到高度优化。此外, r 和 k 作为输入参数,与 n 不存在线性增长关系,这意味着对于长时间序列, r 和 k 也可视作常量。因此,SEDIS的总时间复杂度是近似线性的。

3 实验与结果分析

实验主要分为3个部分:1)验证SEDIS所得的物理状态的可解释性;2)将SEDIS在时间序列分割和状态识别两阶段的准确率和运行效率与其他算法进行对比;3)验证SEDIS对于相关参数的鲁棒性。

3.1 实验设置

SEDIS使用Matlab实现。实验运行于配置Intel Core i5处理器、1.4 GHz主频、带有8 GB内存的MAC笔记本电脑上。

实验数据集主要细分为以下2类:1)Barbet、Fetal、SP02和ECG数据集,这类数据集包括状态变化更少且长度更短的时间序列,选自文献[18]中的32个数据集,其中涵盖医疗、生物、工业等领域相关数据,以证明SEDIS的通用性;2)PAMAP数据集,这类数据集包括状态更多且长度更长的时间序列,选自PAMAP运动数据集^[23-24],是传感器采集的监测数据,反映传感器佩戴者进行的有氧运动种类,如慢走、跑步、踢足球等。5个数据集的具体信息如表1所示。

表1 数据集信息统计

Table 1 Data set information statistics

数据集	长度	分段个数	状态个数
PAMAP	138 876	11	5
Barbet	4 700	6	3
Fetal	18 000	6	3
SP02	17 521	4	2
ECG	17 521	4	2

SEDIS主要涉及参考子序列长度 l 、参考子序列个数 r 以及参考子序列最相似子序列个数 k 这3个参数。由于SEDIS基于子序列相似性,因此 l 的推荐值是代表性子序列的长度。例如,对于心电图(Electrocardiogram, ECG)序列, l 可设置成一次心跳的时长。借助该推论提出一种自动化设置参考子序列长度的方法。具体地,利用快速傅里叶变换将时间序列从时域映射到频域,此时能量最大的频率 f 即主导频率,频率的倒数为代表性子序列的周期 $l = \frac{1}{f}$ 。

根据此规则,对于5个数据集, l 分别取81、20、24、58、13, r 统一取100, k 分别取 $n/600$ 、 $n/100$ 、 $n/200$ 、 $n/100$ 、 $n/100$,其中 n 表示时间序列长度。

实验选取基于子序列相似性的FLUSS算法^[15](仅支持时间序列的分割问题)、基于斜率的pHMM算法^[5]、基于均值和标准差的AutoPlait算法^[7]与SEDIS进行性能对比。FLUSS基于子序列相似性,需要输入子序列长度,为保证一致性将其设置成SEDIS使用的参考子序列的长度 l 。pHMM通过 ε_r 和 ε_s 两个参数限制使用线段拟合原始序列以及构建线段聚类结果产生的误差,在实验中通过调整这两个参数来获得最佳识别结果。AutoPlait是全自动算法,不涉及参数设置。

3.2 算法可解释性验证

在SP02数据集上对SEDIS所得的物理状态进行可解释性验证。SP02数据集由传感器采集的人体血氧饱和度数据组成,血压在一定程度上会对该数据产生影响。为便于展示,对该时间序列进行降采样处理,将长度缩减至350。由图1(a)可以看出,时间序列中存

在2种不同的状态,两者对应的代表性子序列的长度和波形均不一致,2种状态分别表示传感器佩戴者的血压维持正常和血压突然骤降。对比图1(b)的原始标签以及图1(c)由SEDIS得到的识别结果可知,SEDIS识别出的物理状态与真实状态基本吻合,证明了通过子序列相似性区分状态是可行的。pHMM和AutoPlait的识别结果如图1(d)和图1(e)所示。由于FLUSS仅支持序列分割,因此不在此进行展示。从选择的特征进行分析,pHMM基于斜率,将上升段和下降段分为2种状态,最终识别出3种状态,但无法从更宏观的角度描述状态,而AutoPlait基于均值和标准差,由于该案例中的2种状态在均值和标准差方面不具备区分性,因此AutoPlait将整个时间序列视为同一状态,识别效果不理想。

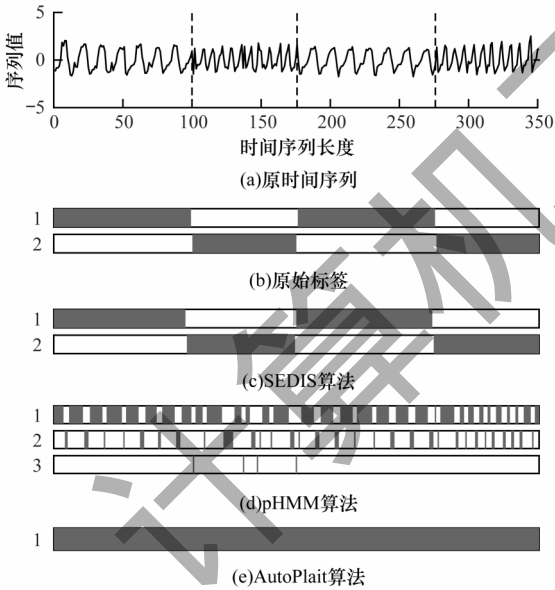


图1 算法可解释性验证结果

Fig.1 Verification results of algorithm interpretability

3.3 算法准确率分析

由于语义挖掘算法在识别出状态的同时也监测出不同状态间的转变点,因此本节将从序列分割和状态识别两个维度对算法准确率进行量化评估。

3.3.1 序列分割准确率

FLUSS 仅能进行序列分割,而其他对比算法忽略了每个分段的状态标识,仅考虑分割点的准确率。本文提出一种新的分割误差指标 s_{sc} ,包括精确率指标 s_{sc_p} 和召回率指标 s_{sc_r} ,其中, s_{sc_p} 为所有预测分割点与其最接近的真值点的距离和, s_{sc_r} 为所有不包含预测分割点的区域的长度和。将整个时间序列划分成多个区域,区域的分界点由两个相邻真值点的中值决定, $s_{sc} = s_{sc_p} + s_{sc_r}$, s_{sc} 越接近于 0,表示分割点越准确。

表2给出了4种算法在5个数据集上的分割误差,其中,×表示该算法未识别出任何有效分割点,意味着将整个时间序列视作同一状态。由表2可以看出,SEDIS

在5个数据集上均具有优异表现。FLUSS同样基于子序列相似性,较优于pHMM和AutoPlait。但由于其难以处理状态多次出现的情况,因此在部分数据集上表现不佳。pHMM更适合区分斜率不同的波段而非复杂波形,因此在结果中出现了大量的分段碎片,导致召回率指标分值较低。AutoPlait难以区分均值和标准差较为一致的状态,仅在两个数据集上找到有效的分割点。由此可见,SEDIS具备更强的通用性。但是,SEDIS未能完全命中真值点的原因在于:时间序列是平滑演进的,分割点附近的子序列相对接近,导致分割点难以精准定位,然而相比于分割点的误判和漏判,准确率方面的误差是可接受的。

表2 4种算法的分割误差比较

Table 2 Comparison of segmentation errors of four algorithms

数据集	SEDIS	FLUSS	pHMM	AutoPlait
PAMAP	0.110 6	0.202 9	62.20	0.403 9
Barbet	0.008 3	0.078 3	24.45	×
Fetal	0.005 1	0.013 9	30.20	0.583 8
SP02	0.004 9	0.079 4	75.29	×
ECG	0.003 0	0.289 4	99.46	×

3.3.2 状态识别准确率

状态识别问题实际上是一个特殊形式的聚类问题,假如将时间序列中的每个时间点视作一个聚类对象,那么目标就是将所有时间点划分成多个不同的类簇,每个类簇反映一个物理状态。本文引入调整兰德系数(Adjusted Rand Index,ARI)^[25]作为状态识别准确率的衡量标准。ARI的取值范围在-1到1之间。ARI越大,表明聚类结果与实际分类情况越吻合。

表3给出了3种算法的状态准确率比较结果,其中×表示该算法未在该数据集上进行状态识别,即将整个时间序列视作一个状态。由表3可以看出,SEDIS在5个数据集上均达到了90%以上的准确率,而pHMM和AutoPlait则表现一般,符合上文的理论分析。

表3 3种算法的状态识别准确率比较

Table 3 Comparison of state recognition accuracy of three algorithms

数据集	SEDIS	pHMM	AutoPlait	%
PAMAP	90.39	2.61	44.10	
Barbet	97.58	9.72	×	
Fetal	98.49	50.25	14.27	
SP02	98.07	0.28	×	
ECG	99.79	19.75	×	

3.4 算法效率分析

基于PAMAP数据集进行算法效率实验,比较不同算法在不同时间序列长度下的运行时间,如图2所示。由图2可以看出,SEDIS的运行效率是近似线性的,优于其他算法。当时间序列长度 n 等于694 380时,SEDIS的运行时间仅为112.73 s。

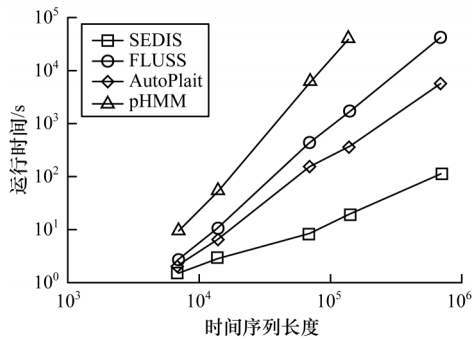


图2 4种算法的运行时间比较

Fig.2 Comparison of the running time of four algorithms

3.5 算法参数设置对状态识别结果的影响

通过实验分析参考子序列长度 l 、参考子序列个数 r 以及参与参考子序列最相似的子序列个数 k 这3个参数在ECG数据集上对于SEDIS状态识别结果的影响程度。假设将子序列长度表示为 l' ,实验中固定其他参数,将子序列长度从 $0.5 \times l'$ 增长至 $4 \times l'$ 。图3(a)给出了不同子序列长度对于状态识别准确率的影响,可以看出SEDIS对于子序列长度的依赖性较弱。图3(b)给出了不同参考子序列个数对于状态识别准确率的影响,可以看出由于SEDIS中基于概率的迭代模式在参考子序列有限的情况下,仍旧具有较为稳定的识别能力。图3(c)给出了与参考子序列最相似的子序列个数,实验从50个子序列增长至400个子序列,可以看出尽管最相似的子序列个数相比其他两个参数对于结果造成的影响更大,但是SEDIS在根据这个指标筛选最相似子序列之后,还会通过基于密度的聚类算法对结果进行再过滤,因此子序列个数的决定性作用被弱化了,使得整体准确率仍保持在98%以上。

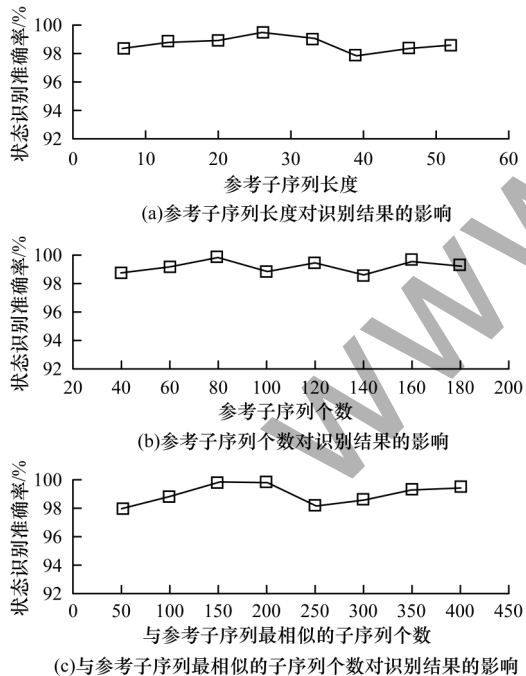


图3 参数设置对状态识别结果的影响

Fig.3 Influence of parameter settings on state recognition results

4 结束语

针对现有时间序列语义挖掘算法缺乏通用性的问题,本文提出一种基于子序列相似性的时间序列语义挖掘算法。通过定义在物理状态层面上的子序列相似性,并结合两级聚类识别时间序列中的潜在状态。引入基于概率的迭代模式,提高算法的识别准确率和运行效率。在真实数据集上的实验结果证明了该算法的有效性和可解释性,并且表明其具有较强的鲁棒性。下一步将针对多维时间序列和流式数据进行扩展,以解决海量数据的实时分析问题,并从语义挖掘的角度出发,将所得状态作为有监督学习样本,进一步执行时间序列的异常检测、预测和分类等流程,实现数据的高效利用。

参考文献

- [1] 黄超. 基于特征分析的金融时间序列挖掘若干关键问题研究[D]. 上海:复旦大学,2005.
HUANG C. Research on several key issues in financial time series mining based on feature analysis[D]. Shanghai: Fudan University, 2005. (in Chinese)
- [2] SAXENA H, ANURAG A V, CHIRAYATH N, et al. Stock prediction using ARMA[J]. International Journal of Engineering and Management Research, 2018, 8(2): 1-4.
- [3] 高飞翔. 心电时间序列的表示方法和相似性度量问题研究[D]. 哈尔滨:哈尔滨工业大学,2014.
GAO F X. Research on representation methods and similarity measures of ECG time series[D]. Harbin: Harbin Institute of Technology, 2014. (in Chinese)
- [4] 史明阳,王鹏,汪卫. 有监督时间序列分割与状态识别算法[J]. 计算机工程, 2020, 46(5): 131-138.
SHI M Y, WANG P, WANG W. Algorithm of supervised time series segmentation and state recognition[J]. Computer Engineering, 2020, 46(5): 131-138. (in Chinese)
- [5] WANG P, WANG H X, WANG W. Finding semantics in time series [C]//Proceedings of 2011 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2011: 385-396.
- [6] EDDY S R. What is a hidden Markov model?[J]. Nature Biotechnology, 2004, 22(10): 1315-1316.
- [7] MATSUBARA Y, SAKURAI Y, FALOUTSOS C. AutoPlait: automatic mining of co-evolving time sequences [C]//Proceedings of 2014 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2014: 193-204.
- [8] KAWABATA K, MATSUBARA Y, SAKURAI Y. StreamScope: automatic pattern discovery over data streams [C]//Proceedings of the 1st International Workshop on Exploiting Artificial Intelligence Techniques for Data Management. New York, USA: ACM Press, 2018: 1-8.
- [9] GRÜNWALD P D. The minimum description length principle[M]. Cambridge, USA: MIT Press, 2007.

- [10] GUI J, ZHENG Z, QIN Z, et al. An approach to extract state information from multivariate time series[J]. *Journal of Computers*, 2020, 31(6): 1-11.
- [11] HONDA T, MATSUBARA Y, NEYAMA R, et al. Multi-aspect mining of complex sensor sequences[C]//*Proceedings of IEEE International Conference on Data Mining*. Washington D. C. , USA: IEEE Press, 2019: 299-308.
- [12] MUEEN A, KEOGH E. Online discovery and maintenance of time series motifs[C]//*Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press, 2010: 1089-1098.
- [13] TOYODA M, SAKURAI Y, ISHIKAWA Y. Pattern discovery in data streams under the time warping distance[J]. *The VLDB Journal*, 2013, 22(3): 295-318.
- [14] 原继东, 王志海, 韩萌. 基于 Shapelet 剪枝和覆盖的时间序列分类算法[J]. *软件学报*, 2015, 26(9): 2311-2325. YUAN J D, WANG Z H, HAN M. Shapelet pruning and Shapelet coverage for time series classification[J]. *Journal of Software*, 2015, 26(9): 2311-2325. (in Chinese)
- [15] GHARGHABI S, DING Y F, YEH C C M, et al. Matrix profile VIII: domain agnostic online semantic segmentation at superhuman performance levels[C]//*Proceedings of IEEE International Conference on Data Mining*. Washington D. C. , USA: IEEE Press, 2017: 117-126.
- [16] GHARGHABI S, YEH C C M, DING Y F, et al. Domain agnostic online semantic segmentation for multi-dimensional time series[J]. *Data Mining and Knowledge Discovery*, 2019, 33(1): 96-130.
- [17] DELDARI S, SMITH D V, SADRI A, et al. ESPRESSO: entropy and shape aware time-series segmentation on for processing heterogeneous sensor data[EB/OL]. [2021-08-11]. <https://arxiv.org/abs/2008.03230v1>.
- [18] YEH C C M, ZHU Y, ULANOVA L, et al. Matrix profile I: all pairs similarity joins for time series; a unifying view that includes motifs, discords and Shapelets[C]//*Proceedings of the 16th International Conference on Data Mining*. Washington D. C. , USA: IEEE Press, 2016: 1317-1322.
- [19] BRACEWELL R N. The Fourier transform and its applications[M]. New York, USA: McGraw-Hill, 1986.
- [20] BIRANT D, KUT A. ST-DBSCAN: an algorithm for clustering spatial-temporal data[J]. *Data & Knowledge Engineering*, 2007, 60(1): 208-221.
- [21] 范子静, 罗泽, 马永征. 一种基于模糊核聚类的谱聚类算法[J]. *计算机工程*, 2017, 43(11): 161-165, 172. FAN Z J, LUO Z, MA Y Z. A spectral clustering algorithm based on fuzzy kernel clustering[J]. *Computer Engineering*, 2017, 43(11): 161-165, 172. (in Chinese)
- [22] HARTIGAN J A, WONG M A. Algorithm AS 136: a k-means clustering algorithm[J]. *Applied Statistics*, 1979, 28(1): 100-108.
- [23] REISS A, STRICKER D. Towards global aerobic activity monitoring [C]//*Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments*. Washington D. C. , USA: IEEE Press, 2011: 1-8.
- [24] REISS A, WEBER M, STRICKER D. Exploring and extending the boundaries of physical activity recognition[C]//*Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*. Washington D. C. , USA: IEEE Press, 2011: 46-50.
- [25] MILLIGAN G W, COOPER M C. A study of the comparability of external criteria for hierarchical cluster analysis[J]. *Multivariate Behavioral Research*, 1986, 21(4): 441-458.

编辑 陆燕菲