Computer Science and Application 计算机科学与应用, 2020, 10(1), 1-10
Published Online January 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.101001

Hans汉斯

Learning from Heterogeneous Temporal Data Based on Electronic Health Records

Min Liang[1], Qian Lu[1], Ningning Li[1], Dong Lin[2], Yuchang Mo[1]

[1]Fujian Province University Key Laboratory of Computational Science School of Mathematical Sciences, Huaqiao University, Quanzhou Fujian
[2]College of Acupuncture, Fujian University of Traditional Chinese Medicine, Fuzhou Fujian
Email: yuchangmo@sina.com

Received: Dec. 9th, 2019; accepted: Dec. 24th, 2019; published: Dec. 31st, 2019

Abstract

Electronic health records contain a large number of longitudinal data, which is valuable for biomedical informatics research. However, standard learning algorithms present challenges due to the complex structure of the data and clinical events that are unevenly distributed over time. Some methods of temporal data modeling depend on extracting single values from time series, which leads to the loss of potentially valuable sequential information. Therefore, how to better explain the temporality of clinical data is still an important research question. In this paper, a new representation of temporal data in electronic health records are studied, which preserves the sequential information that can be processed directly by the standard machine learning algorithms. The research method based on time-series data symbol representation has many different ways. Empirical studies using clinically measured datasets in the real-life database of electronic health records have shown that using distance metrics for random subsequences significantly improves predictive performance compared to the use of original sequences or clustering sequences. The representation method proposed in this paper better explains the temporality of clinical events and is key to the prediction task in the biomedical domain.

Keywords

Electronic Health Record, Random Subsequences, Clustering Sequences, Machine Learning

基于电子健康档案中异构时态数据的学习

梁　敏[1]，陆　迁[1]，李宁宁[1]，林　栋[2]，莫毓昌[1]

[1]华侨大学数学科学学院计算科学福建省高校重点实验室，福建　泉州
[2]福建中医药大学针灸学院，福建　福州
Email: yuchangmo@sina.com

文章引用:梁敏，陆迁，李宁宁，林栋，莫毓昌．基于电子健康档案中异构时态数据的学习[J]．计算机科学与应用，2020, 10(1): 1-10. DOI: 10.12677/csa.2020.101001

收稿日期：2019年12月9日；录用日期：2019年12月24日；发布日期：2019年12月31日

摘　要

电子健康档案包含大量的纵向数据，对于生物医学信息学研究很有价值。然而，由于数据的复杂结构，包括随时间不均匀分布的临床事件，对标准学习算法提出了挑战。时态数据建模的一些方法依赖于从时间序列中提取单一值，导致有潜在价值时序信息的丢失。因此，如何更好地解释临床数据的时效性，仍然是一个重要的研究问题。本文研究了电子健康档案中时态数据新的表示方法，这些表示保留了时序信息，并且可以由标准机器学习算法直接处理。基于时间序列数据符号化表示的研究方法有多种不同的方式。使用电子健康档案真实数据库中临床测量的数据集的实证研究结果表明，相比使用原始序列或聚类序列，对随机子序列使用距离度量显著提高了预测性能。本文提出的表示方法更好地解释了临床事件的时效性，对于生物医学领域的预测任务十分关键。

关键词

电子健康档案，随机子序列，聚类序列，机器学习

1. 引言

电子健康档案(Electronic Health Records, EHRs)包含日常临床活动中获得的大量纵向数据。在 EHRs 中可以获得各种各样的数据类型，系统地收集 EHRs 数据为通过增强错误检测、提高依从性和降低成本来改善临床护理提供了可能性，质量测量、公共卫生监测和患者获得健康状况的数据是 EHR 系统最直接的好处[1]。EHRs 的使用促进了生物医学信息学从个体层面向人群层面转变的机会，这在临床和转化研究中受到越来越多的关注[2]。

与传统的基于叙述的健康档案相比，EHRs 中综合病人病史构成了表型特征的可计算集合，这在很大程度上促进了人口层面的知识发现。同时使用 EHRs 使得数据的异构类型提供了患者的整体视角，其中每位患者的唯一 ID 连接来自不同临床科室的数据，随时间收集的临床数据为患者提供了临床事件的轨迹，因此可以进行纵向分析，这对于生物医学信息学研究很有价值。

目前，EHR 数据的分析主要分为四类：共病性、患者聚类、机器学习和队列查询[3]。其中，机器学习作为一种从大量 EHR 数据中获取有用信息的方法受到了广泛的关注。机器学习算法利用统计理论从训练数据中建立计算模型，并做出可应用于测试数据的推论。通常要求训练数据和测试数据都采用特定的格式，即表格格式，其中对象或示例构成行，描述这些行的特性或属性构成列。但在不损失关键信息的条件下，EHR 数据很少直接符合这种格式，这主要是由于纵向观测的普遍性，即对象不是由具有单个值的一个特征来描述，而是由一段时间内的一系列值来描述。图 1 说明了将 EHR 数据转换成表格格式的问题。通常情况下，一位患者在一定时间内连续不断地报道某些临床测量结果，因此，这些临床测量是结果数据表的相应单元中的一个时间序列。此外，由于每位患者的轨迹都是唯一的，因此这些时间序列通常具有不同的长度，并且以不规则的时间间隔进行测量。标准机器学习算法不能直接从这种复杂的数据

表中构建预测模型，原因主要有两方面：1) 大多数学习算法只能处理包含单个数值或分类值的特征；2) 如果没有参考点，例如给定的查询，大多数学习算法很难比较不同对象的相似程度。因此解决由不同长度和以不规则间隔测量的异构时间序列所增加的复杂性问题十分重要。

如图 1 所示，将原始 EHR 数据转换为用标准机器学习算法直接处理的表格格式时，问题归结为生成表示时间序列的特征。对这一方向已经进行了许多研究[4] [5] [6]。然而，与以前的研究不同，本研究试图处理如图 1 所示的数据表，其中包含从结构丰富的 EHR 数据中提取用于分类的不同长度的时间序列。本研究以检测不良药物事件(ADES)为研究对象，这是一个重大的公共卫生问题。近年来，EHR 数据已成为药品安全监测的宝贵资源。EHRs 不仅具有传统数据来源的优势，而且还提供了与不患有 ADES 的患者形成对照组的可能性。后者尤为重要，因为它允许关联估计，并为有监督的机器学习提供类标签。挖掘用于 ADE 检测的结构化和非结构化 EHR 数据的研究尚处于起步阶段[7] [8]。在 EHRs 中用于 ADE 检测的大多数方法没有考虑到临床事件的时效性，但这对预测任务至关重要。

P1	2013-09-12 10:29:41	M1	0.3
P1	2013-09-13 09:54:20	M1	0.5
P1	2013-09-24 18:46:23	M2	2.8
P1	2013-09-31 11:45:29	M1	0.2
…			
P2	2014-02-15 22:17:24	M1	1.2
P2	2014-02-16 09:32:14	M3	40
P3	2010-11-23 15:14:22	M3	36
…			

⇩

ID	c	M1	M2	M3	…
P1	1	(0.3,0.5,0.2,0.15)	(2.8)	NA	…
P2	0	(1.2)	(3.3,2.5,4.1)	(40)	…
P3	1	(1.3,0.9,1.2)	NA	(36,41,33,40,37)	…
…	…	…	…	…	…

Figure 1. Illustration of the complexity when extracting EHR data
图 1. EHR 数据转换成表格格式示意图

2. 相关工作

为了分析时间序列数据，SAX 表示已广泛用于单变量和多变量时间序列[9] [10]。与本研究中描述的问题密切相关的是，一些研究使用 SAX 表示从时间序列数据中创建特征。为了将时间序列数据转换成表格格式，Hills 等[11]提出了一种单扫描 shapelet 变换算法，该算法从一组单变量时间序列中寻找最佳 k 个 shapelet，然后将原始时间序列数据转换为具有 k 个特征的数据集，每个特征都是该序列与一个 shapelet 之间的距离。这样，任何标准学习算法都可以处理变换后的数据集。Gordon 等[12]引入了 shapelet 采样算法，用于基于 shapelet 分类树的快速计算，对随机子序列进行采样和评估，并计算信息增益。Karlsson 等[13]提出了随机 shapelet 森林算法，该算法在生成森林中的树时，生成多个 shapelets，并在每个节点上选择最好的一个。该方法在保持预测性能的同时，显着地降低了计算成本。

受前面这些方法的启发，我们提出了基于随机子序列的方法处理图 1 所示的数据表。首先应用 SAX 表示从一组时间序列中获得对应于每个特征的符号序列；然后，在每个特征下，随机地生成多个子序列，从中选出信息量最丰富的子序列；最后，将每个时间序列替换为与所选子序列的编辑距离。此时，任何

类型的分类器都可以处理变换后的数据集，生成预测模型受前面这些方法的启发，我们提出了基于随机子序列的方法处理图 1 所示的数据表。首先应用 SAX 表示从一组时间序列中获得对应于每个特征的符号序列；然后，在每个特征下，随机地生成多个子序列，从中选出信息量最丰富的子序列；最后，将每个时间序列替换为与所选子序列的编辑距离。此时，任何类型的分类器都可以处理变换后的数据集，生成预测模型。

3. 基本概念

时间序列 $T=\{t_1,\cdots,t_n\},t_i\in R,\forall i\in[1,n]$ 是在一个时间周期内进行的 n 个度量的有序集合，通常对应于一个时间演化变量。给出固定参数 w，时间序列 T 可以由实向量 $\overline{T}=\{\overline{t_1},\cdots,\overline{t_w}\}$ 表示在 w 维空间中，其中 $\overline{T}$ 的第 i 个元素计算如下：

$$\overline{t_i}=\frac{w}{n}\sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i}t_j \tag{1}$$

通过将每个时间序列分割成大小相等的 w 段，原始时间序列表示从 n 维降到 w 维。每段被分配一个新的值，该值对应于属于该段内的序列的平均值。这种表示被称为分段聚合近似表示(Piecewise Aggregate Approximation, PAA) [14]。

给定时间序列 T，应用 PAA 获得 $\overline{T}$，通过将 $\overline{T}$ 的每个值映射到使用高斯分布定义的离散符号获得 $\overline{T}$ 的离散表示。更确切的说，定义一组 $\alpha-1$ 个分割点 $B=\{\beta_1,\cdots,\beta_{\alpha-1}\}$，其中 α 是字母表大小，使得高斯正态曲线下每对 (β_i,β_{i+1}) 对应的面积都等于 $1/\alpha$。为了完整性，假设 $\beta_0=-\infty,\beta_\alpha=\infty$。因此，给定所需的字母表大小 α，通过查阅统计表可以很容易地定义分割点。

得到分割点后，$\overline{T}$ 中的 PAA 系数将映射得到符号 $\hat{\overline{T}}$ 序列，具体如下：低于第一个分割点的所有系数都映射到第一个字母表符号，例如 a；将第一个和第二个分割点之间的下一组系数值映射到第二个符号，例如 b；以此类推。由此产生的离散序列表示称为符号聚合近似表示(Symbolic Aggregate Approximation, SAX) [15] [16]。

对于一个长度为 w 的离散事件序列 X，X 的子序列 S 被定义为 X 中连续符号长度为 l 的抽样，使得 $l\le w$，即 $S=\{x_q,\cdots,x_{q+l-1}\},l\le q\le w-l+1$。设 $D(\cdot)$ 是两个长度相同的离散事件序列的距离函数。给定长度为 w 的目标序列 X 和长度为 l 的序列 S，$l\le w$，S 到 X 的距离函数 $D(\cdot)$ 定义如下：

$$Dist(S,X)=\min_{\forall S'\subseteq,|S'|=l}\{D(S',S)\} \tag{2}$$

其中 S' 是长度为 l 的时间序列 X 的子序列。$D(\cdot)$ 可以是字符串匹配的任意距离函数，本研究使用编辑距离[17]，也被称为 Levenshtein 距离。

使用上述距离函数，定义随机序列子序列(sequence shapelets)的概念。考虑由 K 类组成的离散序列数据集 D，设 $P(C_i)$ 为属于类 C_i 的序列的比例，$\forall i\in[1,k]$。D 的熵定义为：

$$I(D)=-\sum_{i=1}^{k}p(C_i)\log(p(C_i)) \tag{3}$$

将 D 划分为 m 个不相交的子集 $D_1,\cdots,D_m$，D 的总熵定义为：

$$\hat{I}(D)=-\sum_{i=1}^{m}\frac{D_i}{|D|}I(D_i) \tag{4}$$

因此，数据集 D 上特定分区策略 sp 的信息增益定义为：

$$Gain(D,sp)=I(D)-\hat{I}(D) \tag{5}$$

基于时间序列子序列[18]的原始定义，我们将随机序列子序列定义为 D 中所有子序列中信息增益最高的离散事件子序列。利用等式(3)~(5)，Shapelet S 的信息增益 $SGain(\cdot)$ 计算如下：

$$SGain(S)=\max_{sp\in D} Gain(D,sp) \tag{6}$$

4. 数据分析方法

本节详细描述所提出的方法。首先介绍从时间序列中生成符号序列，然后介绍使用生成的符号序列作为特征进行分类的四种策略：1) 原始序列；2) 聚类序列；3) 随机子序列；4) 随机动态子序列。

4.1. 序列生成

为了使时间序列更容易由分类器直接处理，我们使用 SAX 表示将每个时间序列转换为字符串(图 2)。在 SAX 表示算法中需要设置两个参数：维数(w)和字母表大小(a)。第一个参数 w 是时间序列被划分为大小相等的分段数目，而 a 是用来映射所有分段中标准化时间序列值的符号数。在本研究中，我们考虑 a 值为 2、3 和 5，2 个符号反映高和低(或，正常和异常)，3 个符号增加一个中间范围，5 个符号考虑一个更精确的系统，其中基本时态信息损失更少。通过将 SAX 表示应用于每个特征，由异构时间序列组成的数据集转换为包含不同长度序列的数据集，可以通过标准分类器直接处理。

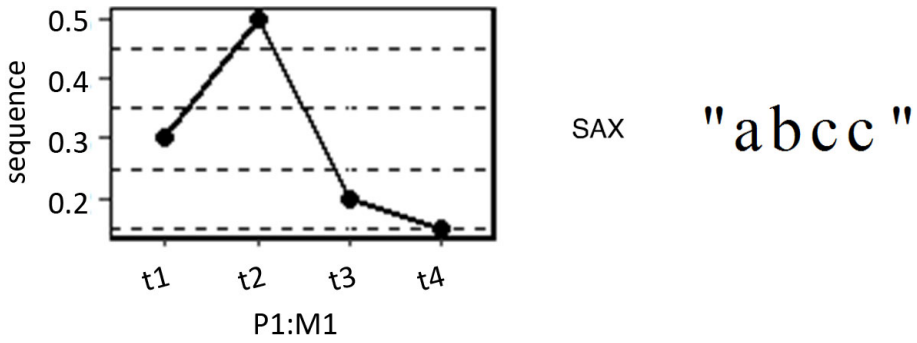

Figure 2. Sequence generation from time series using the SAX representation
图 2. 使用 SAX 表示从时间序列生成序列

4.2. 序列聚类

通过将上述序列生成方法应用于每个时间序列，得到的数据集包含不同长度的序列作为特征。当同一特征中的大多数序列是不同的时，直接使用它们进行分类可能不会产生有意义的结果，因为无法在测试集中的序列与训练集中的序列之间找到匹配。为了减少一个特征中的序列多样性，一种解决方案是将它们进行分组，这样具有相似符号特征和/或符号数目的序列就会分为同一类别。为此，我们使用分区聚类，更具体地说是围绕中心点的划分(PAM)算法[19]对原始序列进行聚类。PAM 算法寻找 k 个有代表性的中心点——类中的对象与所属类中其他所有对象之间的平均相异度度量最小——通过将每个对象分配到与其最近的中心点来构建 k 个类别。在本研究中，对象是通过 SAX 表示生成的序列。每个序列代表一个临床测量的记录，即在规定的时间段内由一个病人得到的变量。它类似于 k-均值聚类算法，但更健壮，因为它允许使用任何类型的距离度量，并最小化每对差异度度量的总和，而不是平方欧几里德距离的总和[20]。将数据集划分为 k 个类别之后，属于同一类别的序列被其中心点代替用于构建预测模型。

Algorithm 1. Random subsequence selection
算法 1. 随机子序列选择

给出一个特征 $f = \{s_1, s_2, \ldots s_n\}$

令 L 为序列 s_i 的最长长度

连接 $s_1, s_2, \ldots s_n$ 构成 S

对 1 到 n_iter 做如下循环

开始索引 k_i：随机数，k_i<序列 S 的长度-L + 1

移动长度 l_i：随机数，$l_i < L$

序列 S 从位置 k_i 移动到 $k_i + l_i$-1 的子序列记为 St_i

f'表示由特征中每个序列与子序列 St_i 的编辑距离组成的特征向量

$IG_i(f')$表示特征向量 f'的信息增益

循环结束

返回信息增益最大所对应的子序列 ST_i

4.3. 随机子序列选择

在本研究中，我们提出了一种减少特征中序列多样性的替代方法。在这里，我们的目标是找到一个子序列，该子序列表示来自同一个类的序列共享的公共符号对齐，并将其与其他类别区分开来。通过使用子序列，具有不同长度的原始序列的问题得到解决，因为长度不同的两个序列只要有相同的代表性子序列，相互之间就是有关的。

对于每个特征，首先使用随机生成的子序列将特征从序列向量转换为原始序列与生成子序列之间的距离向量。这种子序列是从该特征中现有的单个元素创建的，其随机长度不超过最长的原始序列。使用基于等式(3)-(5)的信息增益(IG) [21]评估所生成的子序列。在所生成的子序列中，选择 IG 最高的子序列，然后将特征转化为与原始序列之间的距离。本研究中使用的距离度量是编辑距离[12]，它根据将一个字符串转换到另一个字符串所需的单个字符编辑－插入、删除或替换的最小数量计算。给定由 n 个不同序列 s 组成的特征 f，算法 1 给出了随机子序列选择的详细实现步骤。在本研究中，n_iter 被设置为 10000，充分覆盖每个特征中给定的不同序列数的潜在子序列。

4.4. 随机动态子序列

通过用可变的字母表代替固定不变的字母表大小扩展了随机子序列模型，因为并不是所有的时间序列都必定遵循相同的变化。在本研究中，我们使用算法为每个临床测量动态地分配最合适的 a 值。通过将不同 a 值(2、3 和 5)的 SAX 表示应用于每个临床测量，然后从使用不同 a 值创建的序列组中选择一个最优的子序列来实现，这种方法称为随机动态子序列。算法 2 给出了详细的算法实现步骤。

Algorithm 2. Random dynamic subsequence
算法 2. 随机动态子序列

给出由 N 个特征 $f_1, f_2, \ldots f_N$ 组成的数据集 D

对 1 到 N 做如下循环

对 α 分别为 2,3,5 做循环

将特征 f_i 转换成 SAX 表示 f_i^{α}

f_i^{α} 根据算法 1 得到子序列 st_i^{α}

Continued

f_i^{α}表示 f_i 中每个序列与 st_i^{α}的编辑距离组成的向量

$IG_{\alpha}(f')$表示特征向量 f_i^{α}的信息增益

循环结束

返回信息增益最大所对应的特征向量 f'_i

循环结束

返回 N 个特征 $f'_1,f'_2,...f'_N$的数据集 D'

5. 实证研究

在本节中，我们通过对药品不良事件检测的实证研究，验证能否通过应用基于随机子序列的方法来提高预测性能。首先介绍数据集，然后对所提出的方法进行一系列的实验验证和评估，最后给出预测性能的评价指标。

5.1. 数据集

对于训练模型，实验使用 Adadelta [22]，预处理后共有 72 个样本，其中阳性样本数 11 个，阴性样本数 61 个。我们以 0.75:0.25 的比例将数据集随机划分为训练和测试集。训练集用于模型拟合的数据样本，测试集用来评估最终模型的泛化能力。

5.2. 实验

在本研究中，设计了一系列的实验来研究所提出的方法在分类任务中处理异构时间序列作为特征的影响。

第一个实验评估三组 9 个模型的预测性能。每组包含 a 值为 2、3 和 5 的 SAX 表示得到的数据集。第一组模型使用 SAX 表示得到的序列作为特征，用原始序列表示；第二组模型使用序列聚类得到的中心点作为特征，用聚类序列表示；第三组以序列与其对应的随机子序列之间的编辑距离作为特征，用随机子序列表示。

第二个实验将使用随机动态子序列算法产生特征的模型与将序列长度作为特征的模型进行比较。序列长度是指序列中符号的数量，不考虑时间序列的序列信息。先前研究认为，在 EHRs 中用于 ADEs 检测的临床测量中，序列长度是最优的将时间序列概括为单个值的表示[23]。

后续实验通过变量重要度分析对随机动态子序列模型进行研究，得到使用不同a值生成的子序列中，最好的代表相应临床测量动态表示的子序列。分析不同 a 值在所有临床测量中的分布，以及在信息最丰富的测量中的分布。

5.3. 评估

采用随机森林算法[24]作为底层分类器评估所提出的方法。由于数据集的分类不平衡，本研究中使用的性能评估指标是 ROC 曲线下的区域(AUC) [25] [26]。ROC 曲线代表敏感性(真阳性率)和 1-特异性(假阳性率)之间的一种权衡，前者衡量有多少阳性被识别为阳性，后者衡量有多少阴性被识别为阳性。分析 ROC 曲线的一个常用方法是计算 AUC。本研究使用的另一个评估指标是 F1 测量。F1 值是精确率和召回率的调和均值，是精确率和召回率的综合评价指标。

5.4. 结果

本节给出实施的一系列实验对应的结果。

Table 1. Results of random forest models using Original Sequence (OS), Sequence Cluster (SC) and Random Subsequence (RS)
表 1. 原始序列(OS)、序列聚类(SC)和随机子序列(RS)的随机森林模型的结果

	F1			AUC		
	α = 2	α = 3	α = 5	α = 2	α = 3	α = 5
OS	0.429	0.444	0.364	0.625	0.643	0.625
SC	0.545	0.533	0.471	0.732	0.750	0.679
RS	0.571	0.571	0.667	0.714	0.786	0.750

5.4.1. 比较原始序列、聚类序列和随机子序列

使用字母表大小 a 分别为 2、3 和 5 的原始序列、聚类序列和随机子序列这三种方法建立了 9 个预测模型。表 1 给出了预测性能结果的总结。可以看出，对于选定的 a，模型的选择对预测性能产生一定的影响。总体来说，使用随机子序列模型可以获得最好的预测性能。然而，对于选定的模型，没有迹象表明一个特定的 a 值在所有模型中是最有效的。因此使用随机动态子序列方法，该方法为数据集中的每个特征寻找最合适的 a 值。

5.4.2. 比较随机动态子序列与序列长度

分别使用随机动态子序列和序列长度创建特征的随机森林模型的结果见表 2。显然，前者的使用明显优于后者，尽管序列长度在以前被认为是同一任务中此类数据的最佳单值表示形式。相比于序列长度，使用随机动态子序列方法时，AUC 提高了 11%，F1 提高了 31%。

根据基尼重要性评分计算每个特征的变量重要度后，可以对所有特征进行相应的排序。由于每个特征只包含由一个 a 值获得的序列，因此可以对所有特征的 a 选择的分布进行总结。图 3 以定量和定性的方式说明了 a 的选择。定量地(左图)显示每个 a 值在所有特征中所占的比例。定性地(右图)，根据变量的重要性得分对 a 进行相应的排序。结果表明，数据集的较多特征都是通过 a 值为 2 的 SAX 表示进行转换的。

Table 2. Results on comparing Sequence Length (SL) with Random Dynamic Subsequence (RDS)
表 2. 序列长度(SL)与随机动态子序列(RDS)的比较结果

	F1	AUC
SL	0.444	0.667
RDS	0.643	0.750

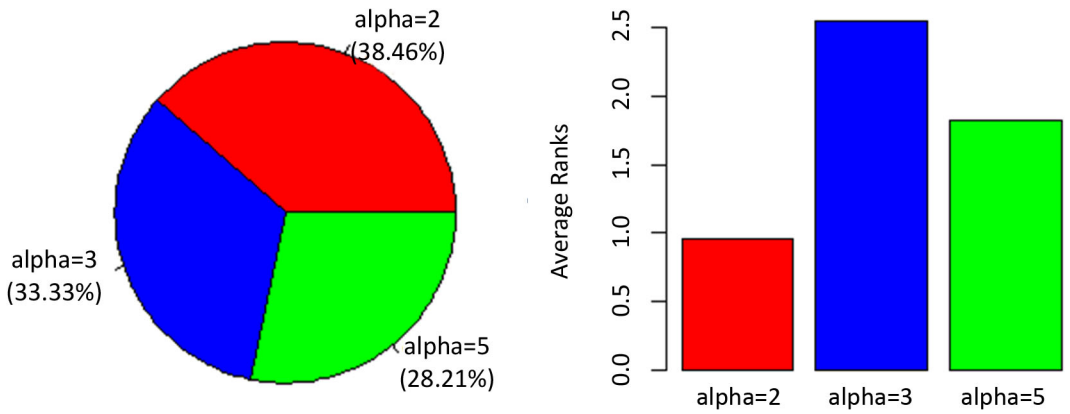

Figure 3. Distribution of a values among all features and the corresponding average ranks according to their variable importance
图 3. a 值在所有特征之间的分布，并根据变量的重要性得分对其进行相应的排序

6. 总结

电子健康档案包含大量纵向、时间戳的临床数据，用机器学习算法处理此类数据通常需要将其转换为表格格式。本研究提出了一种基于子序列的时间序列符号化表示方法。该方法允许直接应用任何标准机器学习算法，同时与基于单一表示的方法相比，它在一定程度上能够获取时序信息，因而显著地提高了预测性能。

基金项目

国家自然科学基金项目(61572442)；福建省高校创新团队发展计划，福建省研究生导师团队，泉州市高层次人才团队项目(2017ZT012)；华侨大学研究生科研创新基金资助项目。

参考文献

[1] Safran, C., Bloomrosen, M., Hammond, W.E., *et al*. (2007) Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, **14**, 1-9. https://doi.org/10.1197/jamia.M2273

[2] Hersh, W.R. (2007) Adding Value to the Electronic Health Record through Secondary Use of Data for Quality Assurance, Research, and Surveillance. *Clinical Pharmacology & Therapeutics*, **81**, 126-128. https://doi.org/10.1038/sj.clpt.6100029

[3] Jensen, P.B., Jensen, L.J. and Brunak, S. (2012) Mining Electronic Health Records: Towards Better Research Applications and Clinical Car. *Nature Reviews Genetics*, **13**, 395-405. https://doi.org/10.1038/nrg3208

[4] Patel, D., Hsu, W. and Lee, M.L. (2008) Mining Relationships among Interval-Based Events for Classification. *Proceedings of the* 2008 *ACM SIGMOD International Conference on Management of Data*, Vancouver, 9-12 June 2008, 393-404. https://doi.org/10.1145/1376616.1376658

[5] Batal, I., Fradkin, D., Harrison, J., Moerchen, F. and Hauskrecht, M. (2012) Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data. *Proceedings of the* 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, 12-16 August 2012, 280-288. https://doi.org/10.1145/2339530.2339578

[6] Zhao, J. and Henriksson, A. (2016) Learning Temporal Weights of Clinical Events Using Variable Importance. *BMC Medical Informatics and Decision Making*, **16**, 71. https://doi.org/10.1186/s12911-016-0311-6

[7] Harpaz, R., Haerian, K., Chase, H.S. and Friedman, C. (2010) Mining Electronic Health Records for Adverse Drug Effects Using Regression Based Methods. *The* 1*st ACM International Health Informatics Symposium*, Arlington, VA, 11-12 November 2010, 100-107. https://doi.org/10.1145/1882992.1883008

[8] Zhao, J., Henriksson, A., Asker, L. and Boström, H. (2015) Predictive Modeling of Structured Electronic Health Records for Adverse Drug Event Detection. *BMC Medical Informatics and Decision Making*, **15**, S1. https://doi.org/10.1186/1472-6947-15-S4-S1

[9] Scheff, J.D., Almon, R.R., Du Bois, D.C., Jusko, W.J. and Androulakis, I.P. (2010) A New Symbolic Representation for the Identification of Informative Genes in Replicated Microarray Experiments. *OMICS*: *A Journal of Integrative Biology*, **14**, 239-248. https://doi.org/10.1089/omi.2010.0005

[10] Siirtola, P., Koskimäki, H., Huikari, V., Laurinen, P. and Röning, J. (2011) Improving the Classification Accuracy of Streaming Data Using Sax Similarity Features. *Pattern Recognition Letters*, **32**, 1659-1668. https://doi.org/10.1016/j.patrec.2011.06.025

[11] Hills, J., Lines, J., Baranauskas, E., Mapp, J. and Bagnall, A. (2014) Classification of Time Series by Shapelet Transformation. *Data Mining and Knowledge Discovery*, **28**, 851-881. https://doi.org/10.1007/s10618-013-0322-1

[12] Gordon, D., Hendler, D. and Rokac, L. (2012) Fast Randomized Model Generation for Shapelet-Based Time Series Classification. *Computer Science*, 1-10.

[13] Karlsson, I., Papapetrou, P. and Boström, H. (2016) Generalized Random Shapelet Forests. *Data Mining and Knowledge Discovery*, **30**, 1053-1085. https://doi.org/10.1007/s10618-016-0473-y

[14] Chakrabarti, K., Keogh, E., Mehrotra, S. and Pazzani, M. (2002) Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. *ACM Transactions on Database Systems*, **27**, 188-228. https://doi.org/10.1145/568518.568520

[15] Lin, J., Keogh, E., Lonardi, S. and Chiu, B. (2003) A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego, CA, 13 June 2003, 2-11. https://doi.org/10.1145/882082.882086

[16] Lin, J., Keogh, E., Wei, L. and Lonardi, S. (2007) Experiencing Sax: A Novel Symbolic Representation of Time Series. *Data Mining and Knowledge Discovery*, **15**, 107-144. https://doi.org/10.1007/s10618-007-0064-z

[17] Levenshtein, V. (1965) Binary Codes Capable of Correcting Spurious Insertions and Deletions of Ones. *Problems of Information Transmission*, **1**, 8-17.

[18] Ye, L. and Keog, E. (2009) Time Series Shapelets: A New Primitive for Data Mining. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 28 June-1 July, 2009, 947-956. https://doi.org/10.1145/1557019.1557122

[19] Kaufman, L. and Rousseeuw, P.J. (1990) Partitioning around Medoids (Program PAM). In: *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 68-125. https://doi.org/10.1002/9780470316801.ch2

[20] Reynolds, A.P., Richards, G., de la Iglesia, B. and Rayward-Smith, V.J. (2006) Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms*, **5**, 475-504. https://doi.org/10.1007/s10852-005-9022-1

[21] Shannon, C.E. (2001) A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**, 3-55. https://doi.org/10.1145/584091.584093

[22] Zeiler, M.D. (2012) ADADELTA: An Adaptive Learning Rate Method. *Computer Science*, 1-6.

[23] Zhao, J., Henriksson, A., Asker, L. and Boström, H. (2014) Detecting Adverse Drug Events with Multiple Representations of Clinical Measurements. 2014 *IEEE International Conference on Bioinformatics and Biomedicine*, Belfast, 2-5 November 2014, 536-543. https://doi.org/10.1109/BIBM.2014.6999216

[24] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[25] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[26] Bradley, A.P. (1997) The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, **30**, 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2