(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113361590 A
(43)申请公布日 2021.09.07

(21)申请号 202110619483.3

(22)申请日 2021.06.03

(71)申请人 电子科技大学
地址 611731 四川省成都市高新区(西区)西源大道2006号

(72)发明人 徐杰 胡堰翔 冯韵霖 方伟政 徐明珠

(74)专利代理机构 北京正华智诚专利代理事务所(普通合伙) 11870
代理人 何凡

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

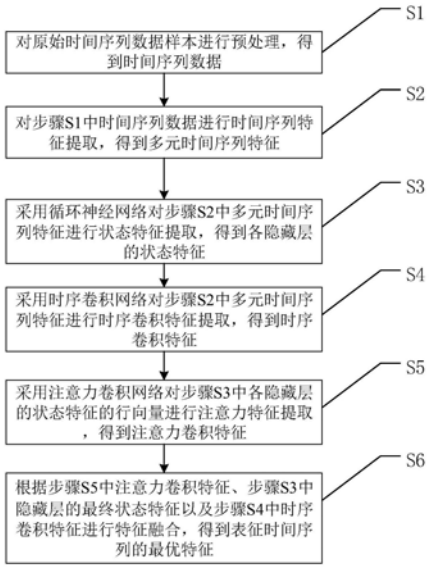
权利要求书2页 说明书10页 附图8页

(54)发明名称

基于多元时间序列的特征融合方法

(57)摘要

本发明公开了一种基于多元时间序列的特征融合方法,通过对包含时序特征的原始数据样本进行预处理,提取处理后的数据特征,构建为二维时间序列数据后,输入时序卷积模块中,得到时序卷积层输出;再将循环神经网络中隐藏层行向量输入注意力卷积模块中,得到最终注意力权重,并根据最终注意力权重计算上下文向量,将得到的上下文向量、时序卷积层输出以及隐藏层的最终状态特征进行融合,得到融合向量,进而可通过构建机器学习验证本发明所得融合特征在多元时间序列数据分析可靠性,具体表现为数据特征提取中良好的泛化性与处理效果,与现有技术相较,本发明能快速获取表征时间序列的最优特征子集,为多元时间序列数据分析提供有效数据支撑。



CN 113361590 A

1.一种基于多元时间序列的特征融合方法,其特征在于,包括以下步骤:

S1、对原始时间序列数据样本进行预处理,得到时间序列数据;

S2、对步骤S1中时间序列数据进行时间序列特征提取,得到多元时间序列特征;

S3、采用循环神经网络对步骤S2中多元时间序列特征进行状态特征提取,得到各隐藏层的状态特征;

S4、采用时序卷积网络对步骤S2中多元时间序列特征进行时序卷积特征提取,得到时序卷积特征;

S5、采用注意力卷积网络对步骤S3中各隐藏层的状态特征的行向量进行注意力特征提取,得到注意力卷积特征;

S6、根据步骤S5中注意力卷积特征、步骤S3中隐藏层的最终状态特征以及步骤S4中时序卷积特征进行特征融合,得到表征时间序列的最优特征。

2.根据权利要求1中所述的所述一种基于多元时间序列的特征融合方法中,所述步骤S1具体包括以下分步骤:

S11、将原始时间序列数据样本进行分类;

S12、将步骤S11中分类后的原始时间序列数据样本按照时间序列进行标定;

S13、对原始时间序列数据样本中异常值与缺失值进行同类均值填充,得到时间序列数据。

3.根据权利要求1中所述的所述一种基于多元时间序列的特征融合方法中,其特征在于,所述步骤S2具体包括以下分步骤:

S21、提取步骤S1预处理后原始数据样本的统计特征;

S22、将步骤S21得到的统计特征整合为二维时间矩阵,得到多元时间序列特征。

4.根据权利要求1中所述的一种基于多元时间序列的特征融合方法中,其特征在于,所述步骤S3具体包括以下分步骤:

S31、构建循环神经网络;

S32、利用循环神经网络中的循环单元对步骤S2中得到的时间序列特征进行递归计算,表示为:

$h^{(t)}=f(s^{(t-1)},X^{(t)},\theta)$

其中,h为循环神经网络RNN的状态,t为时间步,s为内部状态,θ为循环单元内部的权重,f为激活函数;

S33、根据步骤S32中递归结果提取特征,得到各隐藏层的状态特征。

5.根据权利要求1中所述的一种基于多元时间序列的特征融合方法中,其特征在于,所述步骤S4具体包括以下分步骤:

S41、构建不使用池化操作的时序卷积网络;

所述时序卷积网络中卷积层由多个卷积核组成,卷积核的宽度、高度分别与时间序列数量、特征数量相同;

S42、利用步骤S41中时序卷积网络的各卷积核对步骤S2得到的时间序列特征提取卷积特征,表示为:

$h_k=ReLU(W_k*X+b_k)$

其中,k为卷积核数,h_k为卷积核输出,ReLU(.)为激活函数,W_k为可学习参数,*为卷积操

作,X为时间序列数据,b_k为卷积核的偏置;

S43、利用步骤S42得到的卷积特征得到时序卷积层特征。

6.根据权利要求1所述的一种基于多元时间序列的特征融合方法中,其特征在于,所述步骤S5具体包括以下分步骤:

S51、构建注意力卷积网络,并根据设置的时间窗口参数提取步骤S3中各隐藏层的状态特征作为注意力机制输入特征;

S52、利用步骤S51中注意力卷积网络对注意力机制输入特征的行向量进行卷积计算,得到注意力卷积结果,表示为:

$$H_{i,j}^{C}=\sum_{l=1}^{w}H_{i,(t-w-1+l)}\times C_{j,T-w+l}$$

其中,$H_{i,j}^{C}$表示隐藏层的状态特征H中第i行与第j个卷积核的卷积结果,$H_{i,(t-w-1+l)}$为隐藏层的状态特征H的第i行,第(t-w-1+1)列,$C_{j,T-w+1}$为第j个卷积核的第(T-w+1)个元素,T为注意力机制中时间步长,w为注意力卷积网络窗口,1为常数;

S53、利用评分函数根据步骤S52中注意力卷积结果计算注意力权重,表示为:

$$f(H_i^C,h_t)=(H_i^C)^{\mathrm{T}}W_a h_t$$

其中,f(.)为Score评分函数,H_i^C是卷积结果H^C的第i行向量,$(H_i^C)^{\mathrm{T}}$为卷积结果H^C的第i行向量的转置,W_a为可学习模型参数,h_t为循环神经网络最终隐藏层状态特征;

S54、利用Sigmoid函数根据步骤S53中注意力权重计算最终注意力权重,表示为:

$$\alpha_i=Sigmoid\left(f\left(H_i^C,h_t\right)\right)$$

其中,α_i为最终注意力权重,Sigmoid(.)为激活函数;

S55、利用步骤S54中最终注意力权重与步骤S52中卷积结果计算上下文向量,表示为:

$$v_t=\sum_{i=1}^{m}\alpha_i H_i^C$$

其中,v_t为上下文向量,m为隐藏层状态特征的维度。

7.根据权利要求6所述的一种基于多元时间序列的特征融合方法中,其特征在于,所述步骤S6中根据步骤S5中注意力卷积特征、步骤S3中隐藏层的最终状态特征以及步骤S4中时序卷积特征进行融合,得到表征时间序列的最优特征,表示为:

$h'_t=W_h h_t+W_v v_t+W_c c_t$

其中,h'_t为表征时间序列的最优特征,W_h、W_v、W_c分别为可学习模型参数。

基于多元时间序列的特征融合方法

技术领域

[0001] 本发明涉及多元时间序列特征融合处理以及机器学习，属于信息处理技术领域，具体涉及一种基于多元时间序列的特征融合方法。

背景技术

[0002] 近年来，随着互联网和物联网技术的快速普及和发展，由大数据驱动的决策发挥着重要作用。依托于传统机器学习算法和深度学习算法，模型从数据的模式中不断迭代和拟合，最终输出预测结果。在这个过程中，数据起着至关重要的作用。多元时间序列数据中，多元指的是多变量，即在某一时刻需要考虑多个变量对结果的影响。而多元时间序列在我们的日常生活中无处不在，例如：股票的价格、高速路的交通流量、太阳能发电厂的输出量、病人患病等等。在多元时间序列的应用中，人们通常可基于历史观察得到的时间序列信息进一步判断新的趋势或潜在的危险事件，比如可以基于几个小时前的车流量数据来预测交通拥堵情况并设计更好的路线计划、基于股市的历史相关数据预测不久之后的价格来获得更大的利润等等。由于时间序列数据是一种自然模式，因此几乎每个需要人类认知的任务中都存在时间序列数据。在过去的二十年中，时间序列数据的分类被认为是数据挖掘中最具挑战的问题之一，不同的时间序列之间可能存在复杂的动态依赖关系，但却难以被捕获和学习。

[0003] 循环神经网络(Recurrent Neural Network,RNN)是一种处理时间序列数据的神经网络，其以序列的演化方向进行递归。RNN中所有循环单元通过链式连接，输入是时间序列数据。RNN拥有很多特性，比如时序数据所要求的记忆性、神经网络所具有的参数权重共享，并且是图灵完备的。权重共享是指在一次迭代中，循环单元运用相同的权重系数计算所有时间步的变量。与前馈神经网络相比，权重共享降低了RNN的参数量，而且可以捕获序列中随时间变化的特征。此外，RNN满足通用近似定理，即可以按任何精度逼近任意非线性函数，且对状态空间的紧致性没有要求，只需保证足够多的非线性节点。但是，RNN输出所利用的信息不够丰富，会影响最终预测结果。RNN中每一个时间步的state由前一个时间步的state和当前的输入计算得到，因此其包含的信息有限。

[0004] 卷积神经网络(Convolutional Neural Network,CNN)是深度学习的典型算法之一，其本质是前馈神经网络，其中包含使用卷积计算的卷积核结构。卷积神经网络一大特性是对数据集的表征能力很强，依靠卷积核的少量参数以及参数共享的性质就能学习到关键的特征。CNN首先在音视频领域发展起来，因为音视频特征的本质是像素级的数据，可以转换为矩阵处理，卷积核的卷积操作在矩阵数据中可以发挥出良好的效果。

[0005] 目前，很多医院都已启用数字化医学管理系统，因此可以利用病患的历史数据来训练模型，但是由于患者数据是复杂的多元时间序列数据，数据的维度高、覆盖面广，数据清洗将是一个比较有挑战性的工作。因此，如何挖掘分析多元时间序列数据已经成为现如今技术突破的重难点。

发明内容

[0006] 针对现有技术中的上述不足，本发明提供了一种基于多元时间序列的特征融合方法。

[0007] 为了达到上述发明目的，本发明采用的技术方案为：

[0008] 基于多元时间序列的特征融合方法，包括以下步骤：

[0009] S1、对原始时间序列数据样本进行预处理，得到时间序列数据；

[0010] S2、对步骤S1中时间序列数据进行时间序列特征提取，得到多元时间序列特征；

[0011] S3、采用循环神经网络对步骤S2中多元时间序列特征进行状态特征提取，得到各隐藏层的状态特征；

[0012] S4、采用时序卷积网络对步骤S2中多元时间序列特征进行时序卷积特征提取，得到时序卷积特征；

[0013] S5、采用注意力卷积网络对步骤S3中各隐藏层的状态特征的行向量进行注意力特征提取，得到注意力卷积特征；

[0014] S6、根据步骤S5中注意力卷积特征、步骤S3中隐藏层的最终状态特征以及步骤S4中时序卷积特征进行特征融合，得到表征时间序列的最优特征。

[0015] 该方案的有益效果为：

[0016] 1、对包含时序特征数据进行特征提取时，同时提取数据的基本信息，信息全面丰富；

[0017] 2、利用卷积网络构建注意力卷积网络以及时序卷积网络，将特征数据进行组合，能快速获取表征时间序列的最优特征子集；

[0018] 3、丰富了原始时序特征数据特征，为多元时间序列数据分析提供有效数据支撑；

[0019] 4、基于本发明融合原始时序特征数据特征信息，构建特征融合的模型验证该方法具有良好的泛化性能与分类效果。

[0020] 进一步地，所述步骤S1具体包括以下分步骤：

[0021] S11、将原始时间序列数据样本进行分类；

[0022] S12、将步骤S11中分类后的原始时间序列数据样本按照时间序列进行标定；

[0023] S13、对原始时间序列数据样本中异常值与缺失值进行同类均值填充，得到时间序列数据。

[0024] 该进一步方案的有益效果为：

[0025] 将得到的原始数据进行预处理，简化后续数据融合过程的数据误差。

[0026] 进一步地，所述步骤S2具体包括以下分步骤：

[0027] S21、提取步骤S1预处理后原始数据样本的统计特征；

[0028] S22、将步骤S21得到的统计特征整合为二维时间矩阵，得到多元时间序列特征。

[0029] 该进一步方案的有益效果为：

[0030] 将数据整理为二维向量，得到时间序列数据，便于数据处理。

[0031] 进一步地，所述步骤S3具体包括以下步骤：

[0032] S31、构建循环神经网络；

[0033] S32、利用循环神经网络中的循环单元对步骤S2中得到的时间序列特征进行递归计算，表示为：

[0034] $h^{(t)}=f(s^{(t-1)},X^{(t)},)$

[0035] 其中,h为循环神经网络RNN的状态,t为时间步,s为内部状态,θ为循环单元内部的权重,f为激活函数;

[0036] S33、根据步骤S32中递归结果提取特征,得到各隐藏层的状态特征。

[0037] 该进一步方案的有益效果为:

[0038] 一方面得到隐藏层状态特征,便于之后得到注意力向量;另一方面得到最终状态,用于与时序卷积特征和注意力特征融合。

[0039] 进一步地,所述步骤S4具体包括以下分步骤:

[0040] S41、构建不使用池化操作的时序卷积网络;

[0041] 所述时序卷积网络中卷积层由多个卷积核组成,卷积核的宽度、高度分别与时间序列数量、特征数量相同;

[0042] S42、利用步骤S41中时序卷积网络的各卷积核对步骤S2得到的时间序列特征提取卷积特征,表示为:

[0043] $h_k=\mathrm{ReLU}(W_k*X+b_k)$

[0044] 其中,k为卷积核数,h_k为卷积核输出,ReLU(.)为激活函数,W_k为可学习参数,*为卷积操作,X为时间序列数据,b_k为卷积核的偏置;

[0045] S43、利用步骤S42得到的卷积特征得到时序卷积特征。

[0046] 该进一步方案的有益效果为:

[0047] 将得到的时间序列数据输入时序卷积模块,进行时序卷积,得到同一个特征在时间维度上的变化信息以及不同特征在同一时刻的交互信息。

[0048] 进一步地,所述步骤S5具体包括以下分步骤:

[0049] S51、构建注意力卷积网络,并根据设置的时间窗口参数提取步骤S3中各隐藏层的状态特征作为注意力机制输入特征;

[0050] S52、利用步骤S51中注意力卷积网络对时间窗口参数提取后注意力机制输入特征的行向量进行卷积计算,得到注意力卷积结果,表示为:

[0051] $$H_{i,j}^{C}=\sum_{l=1}^{w}H_{i,(t-w-1+l)}\times C_{j,T-w+l}$$

[0052] 其中,$H_{i,j}^{C}$表示隐藏层的状态特征H中第i行与第j个卷积核的卷积结果,$H_{i,(t-w-1+l)}$为隐藏层的状态特征H的第i行,第(t-w-1+l)列,$C_{j,T-w+l}$为第j个卷积核的第(T-w+l)个元素,T为注意力机制中时间步长,w为注意力卷积网络窗口,l为常数;

[0053] S53、利用评分函数根据步骤S52中注意力卷积结果计算注意力权重,表示为:

[0054] $$f(H_i^C,h_t)=(H_i^C)^{\mathrm{T}}W_a h_t$$

[0055] 其中,f(.)为Score评分函数,H_i^C是卷积结果H^C的第i行向量,$(H_i^C)^{\mathrm{T}}$为卷积结果H^C的第i行向量的转置,W_a为可学习模型参数,h_t为循环神经网络最终隐藏层状态特征;

[0056] S54、利用Sigmoid函数根据步骤S53中注意力权重计算最终注意力权重,表示为:

[0057] $$\alpha_i = Sigmoid\left(f\left(H_i^C, h_t\right)\right)$$

[0058] 其中，α_i为最终注意力权重，Sigmoid(.)为激活函数；

[0059] S55、利用步骤S54中最终注意力权重与步骤S52中卷积结果计算上下文向量，表示为：

[0060] $$v_t = \sum_{i=1}^{m} \alpha_i H_i^C$$

[0061] 其中，v_t为上下文向量，m为隐藏层状态特征的维度。

[0062] 该进一步方案的有益效果为：

[0063] 对循环神经网络RNN中隐藏层状态特征的行向量进行一维卷积得到注意力特征，目的是提取隐藏层状态特征中变量在时间维度上的变化模式。

[0064] 进一步地，所述步骤S6中根据步骤S5中注意力卷积特征、步骤S3中隐藏层的最终状态特征以及步骤S4中时序卷积特征进行融合，得到表征时间序列的最优特征，表示为：

[0065] $h'_t = W_h h_t + W_v v_t + W_c c_t$

[0066] 其中，h'_t为表征时间序列的最优特征，W_h、W_v、W_c分别为可学习模型参数。

[0067] 该进一步方案的有益效果为：

[0068] 将时序卷积特征、注意力卷积特征以及循环神经网络隐藏层的最终状态特征输出进行融合，丰富了数据特征，可为构建多元时间序列预测模型提供数据支撑。

附图说明

[0069] 图1为本发明提供的基于多元时间序列的特征融合方法的步骤示意图；

[0070] 图2为本发明提供的基于多元时间序列的特征融合方法的结构示意图；

[0071] 图3为本发明中步骤S1的步骤示意图；

[0072] 图4为本发明中步骤S2的步骤示意图；

[0073] 图5为本发明中发生AKI的病人住院时间分布；

[0074] 图6为本发明中住院少于7日的AKI病人住院时间分布；

[0075] 图7为本发明中步骤S3的步骤示意图；

[0076] 图8为本发明中循环神经网络RNN的结构示意图；

[0077] 图9为本发明中步骤S4的步骤示意图；

[0078] 图10为本发明中时序卷积模块的结构示意图；

[0079] 图11为本发明中步骤S5的步骤示意图；

[0080] 图12为本发明中注意力卷积模块的结构示意图；

[0081] 图13为本发明中本地模型对比结果柱形图；

[0082] 图14为本发明中主流模型对比结果柱形图。

具体实施方式

[0083] 下面对本发明的具体实施方式进行描述，以便于本技术领域的技术人员理解本发明，但应该清楚，本发明不限于具体实施方式的范围，对本技术领域的普通技术人员来讲，

只要各种变化在所附的权利要求限定和确定的本发明的精神和范围内，这些变化是显而易见的，一切利用本发明构思的发明创造均在保护之列。

[0084] 如图1、图2所示，本发明提供的基于多元时间序列的特征融合方法，包括以下步骤S1至步骤S6：

[0085] S1、对原始时间序列数据样本进行预处理，得到时间序列数据；

[0086] 本实施例中，如图3所示，步骤S1具体包括以下分步骤：

[0087] S11、将原始时间序列数据样本进行分类；

[0088] 实际中，多元时间序列数据在生活中随处可见，医院住院患者的病理数据则为典型的时间序列数据，本发明以某医院2019年全年关于急性肾损伤(AKI)数据为例，将包含时序特征的原始数据样本进行归纳整理，整理为统一格式的数据集，并按照检测频率将指标进行排序，选择检测频率最高的50种指标作为原始数据本的参考指标，可将原始数据的参考指标划分为血常规、肝功能和肾功能三大类。

[0089] S12、将步骤S11中分类后的原始时间序列数据样本按照时间序列进行标定；

[0090] 实际中，将每个原始数据集按照多元时间序列进行判断，以7天内比基线升高1.5倍以上为AKI判断依据，打上标签；除了病历数据中的日常检测指标以外，还将病人的基本信息纳入数据整理范围，由于基本信息中的性别与民族为类别特征，因此需要通过独热编码将类别特征转换为数字特征，进而得到时间序列数据，其中，独热编码将数据中离散特征扩展到了欧式空间，而大部分算法是基于向量空间的度量来计算的，此时离散特征对应为欧式空间中的一个位置，会让特征之间的距离计算具有更高的合理性。

[0091] S13、对原始时间序列数据样本中异常值与缺失值进行同类均值填充，得到时间序列数据。

[0092] 实际中，原始时间序列数据样本存在很多异常的数据，比如数字中夹杂着特殊符号，可能是医生在输入数据时粗心造成的，本发明中将这些异常值作为缺失值来对待，采用同类均值填充方法将基础数据中异常值和缺失值整合为时间序列形式，即将患者的制表时间点排列出来，组成序列；但是由于患者数据并非为每个时间点都有检测目标，因此会出现大量缺失，则将这些数据中上一时间点的指标数据填充到缺失值中处理。

[0093] S2、提取步骤S1预处理后原始数据的时间序列特征，得到时间序列数据；

[0094] 本实施例中，如图4所示，步骤S2具体包括以下分步骤：

[0095] S21、提取步骤S1预处理后原始数据的统计特征；

[0096] 实际中，在对于能天然捕捉时间序列依赖关系的模型比如RNN、LSTM等，需要把数据整合为二维矩阵即可，不需要添加额外的时间特征，故此本发明将提取AKI数据中的统计特征，时间范围基于病历数据中病人从住院起到发生AKI的前一天，具体统计量包括最大值，最小值，均值以及标准差；提取完统计特征后，会出现少量的统计特征缺少，本发明中，直接将该部分缺失统计特征的参数直接删除，因为出现统计特征缺失，则说明该病人的病历数据在该时间范围的特征全部缺失，在提取数据统计过程中，只有少部分的原始数据缺失统计特征，因此删除该部分的数据对数据特征处理不会出现较大干扰；

[0097] S22、将步骤S21得到的统计特征整合为二维向量，得到时间序列数据。

[0098] 实际中，将统计特征整合为二维时间矩阵的形式，需要考虑时间窗的大小，即使用多少天的数据作为输入，由于AKI的判断依据是肌酐水平在七天内是否升高1.5倍以上，因

此七天是一个比较理想的时间窗口大小，但是原始数据的整理后发现有25％的病人在住院的第四天内发生AKI，有50％的病人在住院的第七天内发生AKI，如图5、图6所示，大部分病人的住院时间是满足不了7天的需求，如果把时间窗大小设置为7天，那么至少有一半原始数据无法使用，由病人住院时间分布可以发现，绝大部分病人住院的时间跨度大于3天，因此将时间窗大小设置为3天。

[0099] S3、采用循环神经网络对步骤S2中多元时间序列特征进行状态特征提取，得到各隐藏层的状态特征；

[0100] 本实施例中，如图7所示，步骤S3具体包括以下分步骤：

[0101] S31、构建循环神经网络；

[0102] S32、利用循环神经网络中的循环单元对步骤S2中得到的时间序列特征进行递归计算，表示为：

[0103] $h^{(t)}=f(s^{(t-1)},X^{(t)},\theta)$

[0104] 其中，h为循环神经网络RNN的隐藏层状态，每一个时间步的计算都将利用隐藏层状态，t为时间步，s为内部状态，由系统状态表示为：s＝s(h,X,y)，X为时间序列特征，θ为循环单元内部的权重，与时间步无关，通过反向传播来更新，f为激活函数；

[0105] S33、根据步骤S32中递归结果提取特征，得到各隐藏层的状态特征。

[0106] 实际中，如图8所示，循环神经网络RNN的核心结构是一个有向图，有向图中以链式连接的元素称为循环单元，给定输入的时间序列特征，循环神经网络RNN的总长度为τ，当前时间步的系统状态将由之前的内部状态得到，而内部状态的计算中又包含有系统状态，所以循环单元的计算是一种递归，循环神经网络RNN中常见的激活函数为Sigmoid函数和双曲正切函数Tanh。

[0107] S4、采用时序卷积网络对步骤S2中多元时间序列特征进行时序卷积特征提取，得到时序卷积特征；

[0108] 本实施例中，如图9所示，步骤S4具体包括以下分步骤：

[0109] S41、构建不使用池化操作的时序卷积网络；

[0110] 所述时序卷积网络中卷积层由多个卷积核组成，卷积核的宽度、高度分别与时间序列数量、特征数量相同；

[0111] 实际中，如图10所示，将时序卷积模块引入到循环神经网络中，通过时间序列数据X，n为特征数量，T为时间序列长度，本发明中T＝3，即有三个时间序列，n为特征数量，包括病人的检测指标、基础信息、用药情况等等，通过引入一个不使用池化操作的卷积网络，提取时间维度上的短期依赖关系和特征间的相互依赖关系，每个卷积核的宽度为w，高度为n，其中w与时间窗参数相等，即卷积核的高度和特征数量相同，每个卷积操作将充分考虑每个特征，将该数据输入到基于循环神经网络的时序卷积模块中得到卷积层输出。

[0112] S42、利用步骤S41中时序卷积网络的各卷积核对步骤S2得到的时间序列特征提取卷积核特征，表示为：

[0113] $h_k=\mathrm{ReLU}(W_k*X+b_k)$

[0114] 其中，k为卷积核数，h_k为卷积核特征，ReLU(.)＝max(0,x)是激活函数为激活函数，W_k为可学习参数，*为卷积操作，X为时间序列数据，b_k为卷积核的偏置；

[0115] S43、利用步骤S42得到的卷积核特征得到时序卷积特征。

[0116] 实际中，时序卷积特征大小为r×1，r为卷积核的数量。

[0117] S5、采用注意力卷积网络对步骤S3中各隐藏层的状态特征的行向量进行注意力特征提取，得到注意力卷积特征；

[0118] 本实施例中，如图11所示，步骤S5具体包括以下分步骤：

[0119] S51、构建注意力卷积网络，并根据设置的时间窗口参数提取步骤S3中各隐藏层的状态特征作为注意力机制输入特征；

[0120] 实际中，数据集表示在时间步i观测到的数据，特征一共有n维。设置一个时间窗口参数大小为w，即以时间窗输入数据作为注意力卷积网络的输入。

[0121] S52、利用步骤S51中注意力卷积网络对注意力机制输入特征的行向量进行卷积计算，得到注意力卷积结果，表示为：

[0122] $$H_{i,j}^{C}=\sum_{l=1}^{w}H_{i,(t-w-1+l)}\times C_{j,T-w+l}$$

[0123] 其中，$H_{i,j}^{C}$表示隐藏层的状态特征H中第i行与第j个卷积核的卷积结果，$H_{i,(t-w-1+l)}$为隐藏层的状态特征H的第i行，第(t-w-1+l)列，$C_{j,T-w+l}$为第j个卷积核的第(T-w+l)个元素，T为注意力机制中时间步长，w为注意力卷积网络窗口，l为常数；

[0124] 实际中，大多数时间序列中使用的经典注意力机制中注意力向量(上下文向量)都是对循环神经网络RNN中所有隐藏层的列向量进行加权求和，但是这种计算方式无法忽略对噪声特征，即对列向量进行加权求和没有实际明确的意义，反而是引入更多的噪声信息，同时经典注意力机制会在所有的时间步上去加权和，进行平均信息处理，因此无法捕捉隐藏层状态特征的时间变化模式。

[0125] 本发明中改进了注意力机制中的计算方法，聚焦于隐藏层状态特征的行向量上，行向量指的是某一变量在所有时间步上的表示，有助于优化噪声带来的干扰，通过对循环神经网络中RNN隐藏层状态特征$H\in\mathbb{R}^{m\times(t-1)}$中的行向量进行卷积计算，具体来说，k个卷积核，T是注意力计算中关注的时间步长，大小与时间窗口参数w相同，经过卷积计算，得到注意力卷积结果$H^{C}\in\mathbb{R}^{m\times k}$，m为隐藏层状态特征的维度，k为卷积个数。

[0126] S53、利用评分函数根据步骤S52中注意力卷积结果计算注意力权重，表示为：

[0127] $$f(H_i^{C},h_t)=(H_i^{C})^{\mathrm{T}}W_a h_t$$

[0128] 其中，f(.)为Score评分函数，定义为$f:\mathbb{R}^{k}\times\mathbb{R}^{m}\mapsto\mathbb{R}$，其中，k为卷积个数，m为隐藏层状态特征的维度，$H_i^{C}$是卷积结果$H^{C}$的第i行向量，$(H_i^{C})^{\mathrm{T}}$为卷积结果$H^{C}$的第i行向量的转置，$W_a$为可学习模型参数，$W_a\in\mathbb{R}^{k\times m}$，$h_t$为循环神经网络最终隐藏层状态特征，$h_t\in\mathbb{R}^{m}$；

[0129] S54、利用激活函数根据步骤S53中注意力权重计算最终注意力权重，表示为：

[0130] $$\alpha_i=Sigmoid\left(f\left(H_i^{C},h_t\right)\right)$$

[0131] 其中，α_i为最终注意力权重，Sigmoid(.)为激活函数；

[0132] 实际中，在典型的注意力机制的计算中，注意力权重是由Softmax函数计算的，而本发明中采用Sigmoid函数来替代Softmax函数，Softmax函数的作用是把注意力的权重进行归一化，当有多个特征对模型预测有帮助时，归一化会降低这些特征的重要程度，而归一化并不是必须的，当不使用归一化时，对模型有帮助的多个变量都会起到相应的作用。

[0133] S55、利用步骤S54中最终注意力权重与步骤S52中卷积结果计算上下文向量，表示为：

[0134] $$v_t = \sum_{i=1}^{m} \alpha_i H_i^C$$

[0135] 其中，v_t为上下文向量，m为隐藏层状态特征的维度。

[0136] 实际中，如图12所示，注意力卷积特征即上下文向量是RNN中所有隐藏层状态特征的列向量的加权和，隐藏层状态特征中每个列向量都代表着一个语义空间，但是这种注意力机制的计算方式无法忽略对预测而言是噪声的特征，即对列向量进行加权求和后并没有明确的意义，反而会引入更多的噪声信息，因此本发明中，提出的注意力计算方法将聚焦在隐藏层状态特征的行向量上，行向量指的是某一个变量在所有时间步上的表示，基于行向量的注意力权重的计算有助于帮助模型选择有利于预测的特征。

[0137] S6、根据步骤S5中注意力卷积特征、步骤S3中隐藏层的最终状态特征以及步骤S4中时序卷积特征进行融合，得到表征时间序列的最优特征。

[0138] 本实施例中，步骤S6中根据据步骤S5中注意力卷积特征、步骤S3中隐藏层的最终状态特征以及步骤S4中时序卷积特征进行融合，得到表征时间序列的最优特征，表示为：

[0139] $h'_t = W_h h_t + W_v v_t + W_c c_t$

[0140] 其中，h'_t为表征时间序列的最优特征，W_h、W_v、W_c分别为可学习模型参数，r为卷积核的数量，m为隐藏层状态特征的维度，k为卷积个数。

[0141] 实际中，注意力卷积特征即为上下文向量。

[0142] 如图13所示，基于本发明融合原始时序特征数据特征信息，构建基于特征融合的模型与逻辑回归、随机森林、三层DNN、基础循环神经网络RNN、融合时间序列特征的循环神经网络RNN、融合注意力特征的循环神经网络RNN以及融合时间序列和注意力特征的循环神经网络RNN等本地模型进行比对评估，判断融合表现的好坏；采用AUC和召回率为评估标准，其中AUC(Area Under Curve)表示ROC曲线下的面积ROC曲线全称为受试者工作特征曲线(Receiver Operating Characteristic Curve)，各模型的AUC和召回率如表1所示，

[0143] 表1本地模型对比分析结果

[0144]

模型	AUC	召回率
逻辑回归模型	0.814	0.732
随机森林模型	0.853	0.716
三层 DNN 模型	0.857	0.804
基础 RNN 模型	0.865	0.827
基于时序卷积的模型	0.881	0.846
融合注意力特征的模型	0.887	0.861
基于特征融合的模型	**0.908**	**0.869**

[0145] 从模型结果可以看到,逻辑回归的表现是最一般的,也是符合情理之中,因为线性模型的表征能力是有限的,在复杂的场景下往往没有非线性模型效果好。随机森林的AUC比逻辑回归提高不少,而召回率却没有逻辑回归的高,说明随机森林根据若干决策树训练后的投票结果,更倾向于谨慎地判断病人会发生AKI,即由KS值确定的阈值更偏向于精确率,因此召回率会降低。三层的DNN模型相比于随机森林,AUC并没有提高很多,但召回率却有了较大幅度的增加,说明全连接神经网络模型倾向于大胆的预测病人会发生AKI,同时也保证了预测的较高准确性,DNN结构虽不复杂,但其学习能力很强大,仅根据时间序列数据的统计量特征就能准确地捕捉到特征间的复杂依赖关系,并预测出较准确的结果。基础RNN与DNN相比,AUC和召回率均有小幅提升,从这个模型开始,输入数据由特征统计量变为了时间窗口为3的时间序列数据,从结果可以发现模型从原始时间序列中可以更好地学习数据中的规律,因为当特征统计量作为模型的输入时,这些数据时经过人为处理过的,会造成大量的信息损失,一方面可以减少噪声信息,但另一方面也会把对模型有帮助的信息过滤掉。具有记忆特性的RNN模型从原始时间序列数据中可以学习到复杂的时间模式,因此表现要优于全连接神经网络。

[0146] 仅把时间序列数据的二维卷积特征与循环神经网络RNN的状态特征融合后,AUC和召回率就有了两个百分点的提升,说明从原始数据样本中提取的特征信息会对模型学习很有帮助,这些信息包括两方面的内容:一是同一特征在时间维度上的变化信息,而是同一个时间步上所有特征间的依赖信息。而基于注意力向量的卷积特征融合模型的召回率提升的更为明显,说明循环神经网络RNN中隐藏层状态特征的卷积提供的信息将有助于提高召回率。循环神经网络RNN隐藏层状态特征中变量在时间维度上的卷积具有比原始数据样本的卷积信息更复杂的高阶特征,正是这些高阶特征促进了召回率的提高。

[0147] 最终本发明提出的融合方法构建的基于特征融合模型的AUC为0.908,召回率为0.869,取得了不错的表现。特别是在AUC方面的提升,将时间序列数据的卷积特征和注意力向量的卷积特征融合后,再加上原本循环神经网络RNN的状态特征,这些融合后的特征提供足够丰富的信息,当然在提供更多有用的信息同时,也会引入更多的噪声信息,需要具有丰富的深度学习模型训练经验,通过增加正则方法和参数调整来使特征融合方法达到最佳效果,在保证训练结果良好的同时,提高特征融合方法的泛化能力。

[0148] 如图14所示,本发明提出的基于特征融合的模型与逻辑回归、随机森林、广义加性模型、梯度提升模型、循环神经网络等主流模型进行对比,可以得知:广义加性模型和随机森林集成的模型AUC为0.8,其只使用了病人手术前的数据,即病人住院前的数据,此模型从

框架到特征都比较简单，适合计算力较小的应用场景；离散时间生存模型和逻辑回归集成的模型AUC仅为0.76，但它使用的病人样本数量仅为2122名病人，适用于少样本的场景；梯度提升模型和决策树集成模型的AUC达到了0.87—0.90，这两个模型使用的训练数据均超过了10万名病人，取得了不错的效果；三层RNN模型的AUC是主流模型中最高的，为0.934，此模型由DeepMind团队提出，是目前为止表现最好的模型。性能优秀代表着该模型使用的数据特征会较为复杂，比如该模型对病人进行了筛选，要求病人在入院前必须有至少一年的电子病例记录，而且将病人五年内的历史信息进行汇总并作为模型的特征。这种病人筛选方式以及特征的提取模式仅能适用于极少数的医院场景，这也是该模型的一个局限性，本发明提供的基于特征融合模型的AUC为0.908，虽然没有DeepMind提出的模型AUC高，但比大部分主流模型的AUC高，其样本特征没有DeepMind模型特征复杂的，可适合与普遍的日常检测数据与基础信息。

[0149] 以上通过建立基于特征融合的模型对进行了详细的说明，验证了本发明提供的一种基于多元时间序列的特征融合办法在提取表征时间序列的最优特征，具有明显的提高，证明了本发明方法的有效性。

[0150] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0151] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0152] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0153] 本发明中应用了具体实施例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

[0154] 本领域的普通技术人员将会意识到，这里所述的实施例是为了帮助读者理解本发明的原理，应被理解为本发明的保护范围并不局限于这样的特别陈述和实施例。本领域的普通技术人员可以根据本发明公开的这些技术启示做出各种不脱离本发明实质的其它各种具体变形和组合，这些变形和组合仍然在本发明的保护范围内。

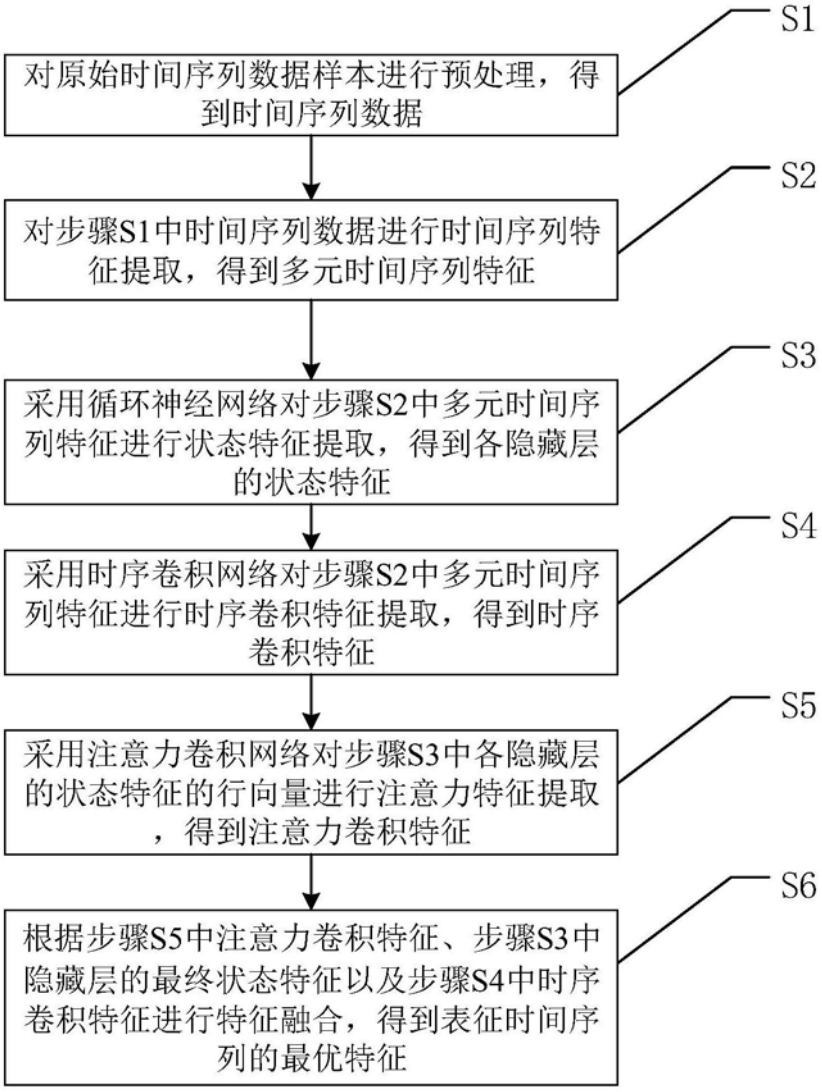

S1
对原始时间序列数据样本进行预处理，得到时间序列数据
S2
对步骤S1中时间序列数据进行时间序列特征提取，得到多元时间序列特征
S3
采用循环神经网络对步骤S2中多元时间序列特征进行状态特征提取，得到各隐藏层的状态特征
S4
采用时序卷积网络对步骤S2中多元时间序列特征进行时序卷积特征提取，得到时序卷积特征
S5
采用注意力卷积网络对步骤S3中各隐藏层的状态特征的行向量进行注意力特征提取，得到注意力卷积特征
S6
根据步骤S5中注意力卷积特征、步骤S3中隐藏层的最终状态特征以及步骤S4中时序卷积特征进行特征融合，得到表征时间序列的最优特征

图1

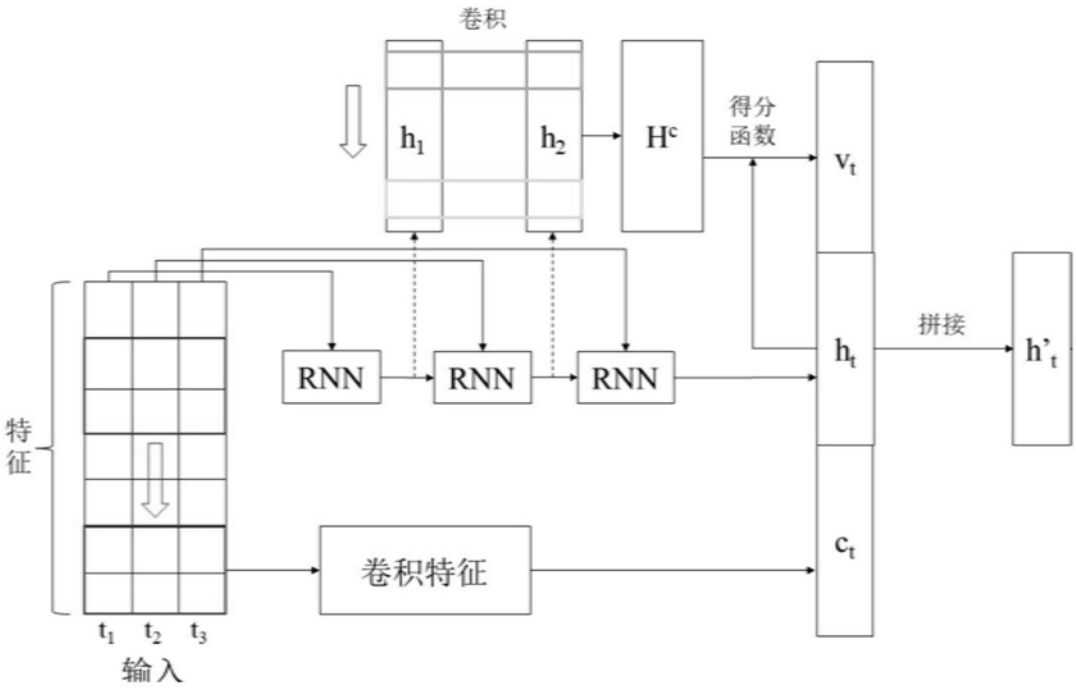

卷积
h_1
h_2
H^c
得分
函数
v_t
RNN
RNN
RNN
h_t
拼接
h'_t
特
征
卷积特征
c_t
t_1 t_2 t_3
输入

图2

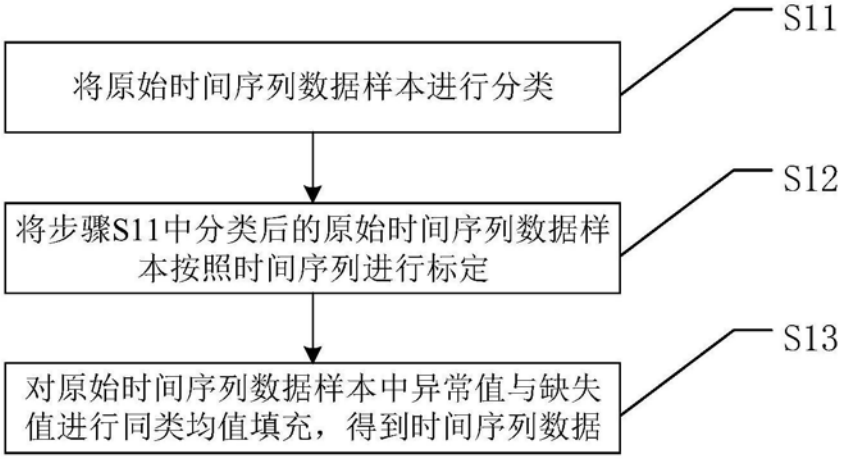

S11
将原始时间序列数据样本进行分类
S12
将步骤S11中分类后的原始时间序列数据样本按照时间序列进行标定
S13
对原始时间序列数据样本中异常值与缺失值进行同类均值填充，得到时间序列数据

图3

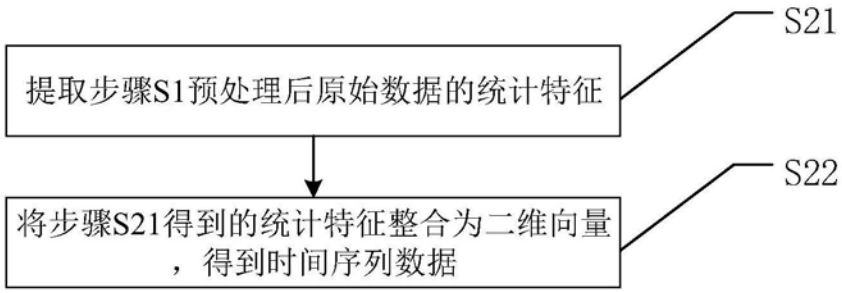

S21
提取步骤S1预处理后原始数据的统计特征
S22
将步骤S21得到的统计特征整合为二维向量，得到时间序列数据

图4

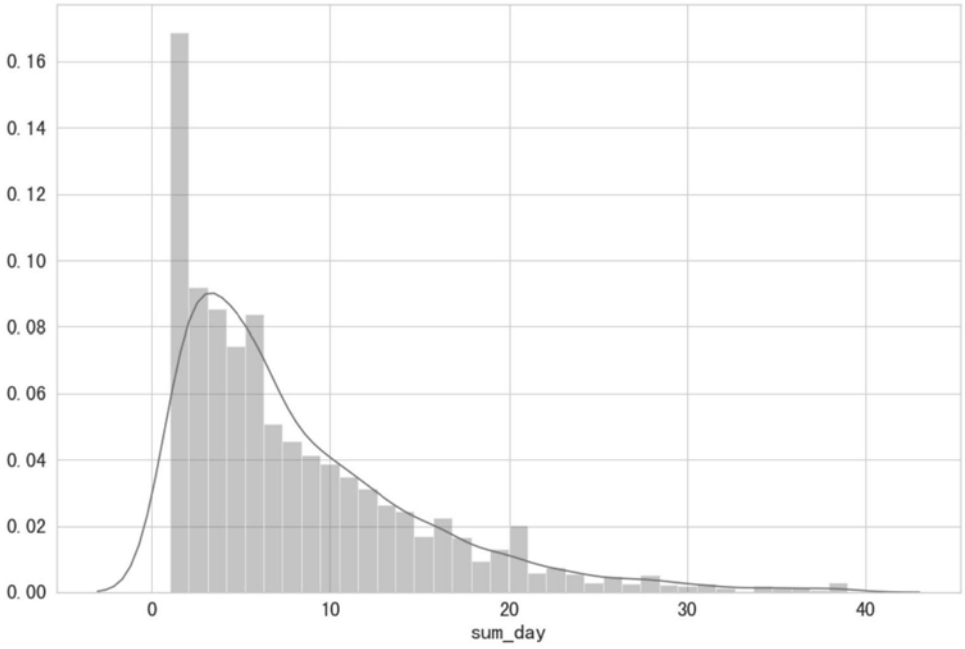

0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
sum_day

图5

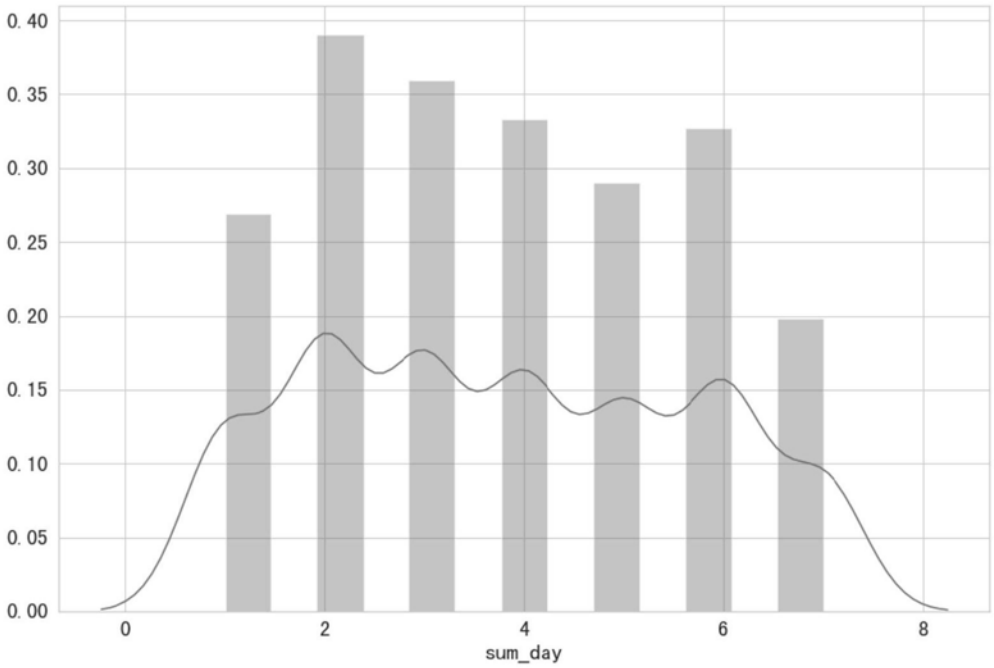

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
sum_day

图6

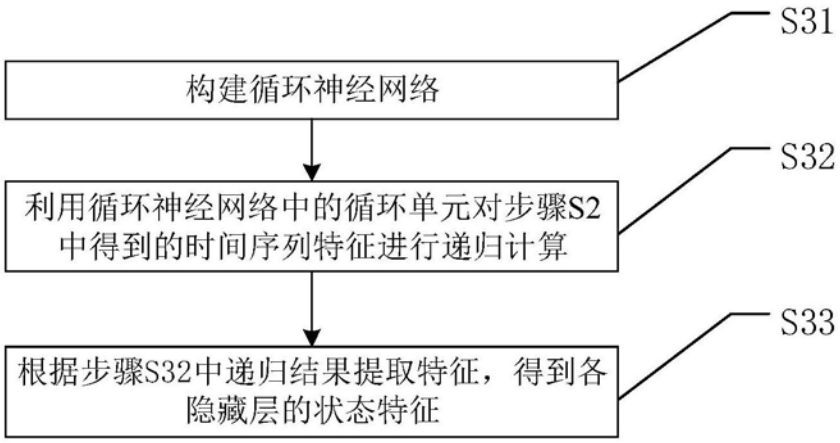

S31
构建循环神经网络
S32
利用循环神经网络中的循环单元对步骤S2中得到的时间序列特征进行递归计算
S33
根据步骤S32中递归结果提取特征，得到各隐藏层的状态特征

图7

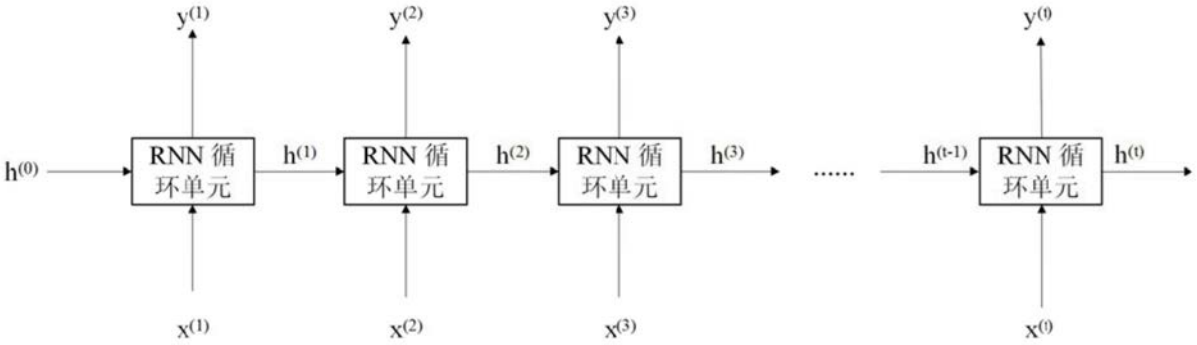

y(1)
y(2)
y(3)
y(t)
h(0)
RNN 循环单元
h(1)
RNN 循环单元
h(2)
RNN 循环单元
h(3)
.....
h(t-1)
RNN 循环单元
h(t)
x(1)
x(2)
x(3)
x(t)

图8

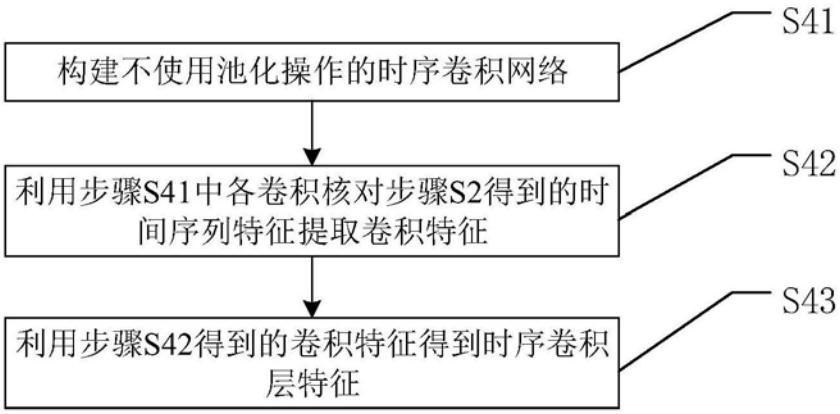

S41
构建不使用池化操作的时序卷积网络
S42
利用步骤S41中各卷积核对步骤S2得到的时间序列特征提取卷积特征
S43
利用步骤S42得到的卷积特征得到时序卷积层特征

图9

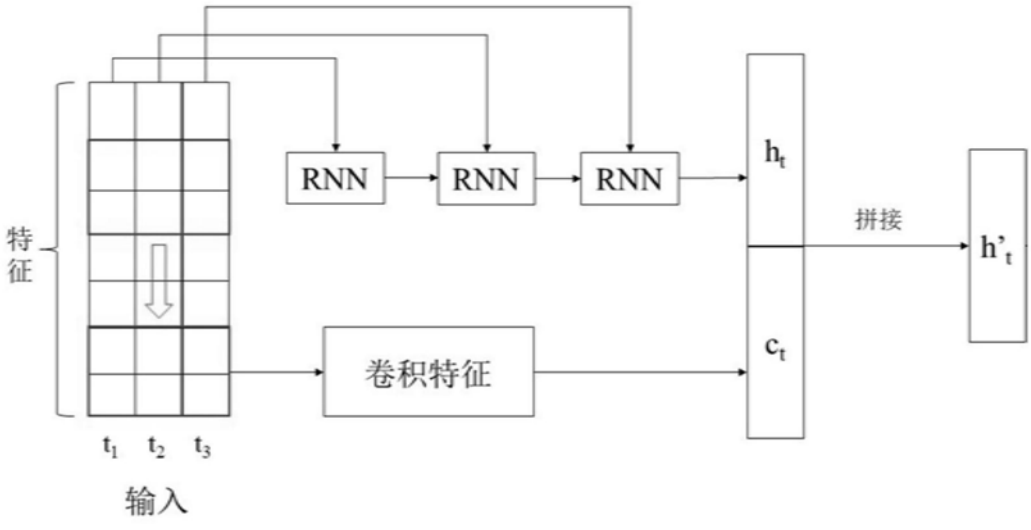

特征
RNN
RNN
RNN
h_t
拼接
h'_t
卷积特征
c_t
t1 t2 t3
输入

图10

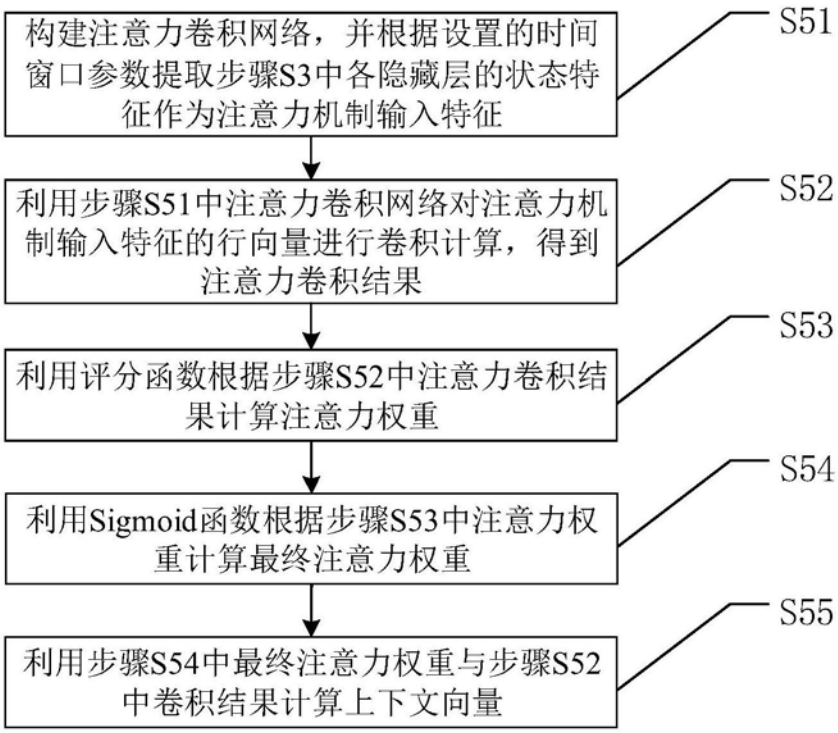

构建注意力卷积网络，并根据设置的时间窗口参数提取步骤S3中各隐藏层的状态特征作为注意力机制输入特征
S51
利用步骤S51中注意力卷积网络对注意力机制输入特征的行向量进行卷积计算，得到注意力卷积结果
S52
利用评分函数根据步骤S52中注意力卷积结果计算注意力权重
S53
利用Sigmoid函数根据步骤S53中注意力权重计算最终注意力权重
S54
利用步骤S54中最终注意力权重与步骤S52中卷积结果计算上下文向量
S55

图11

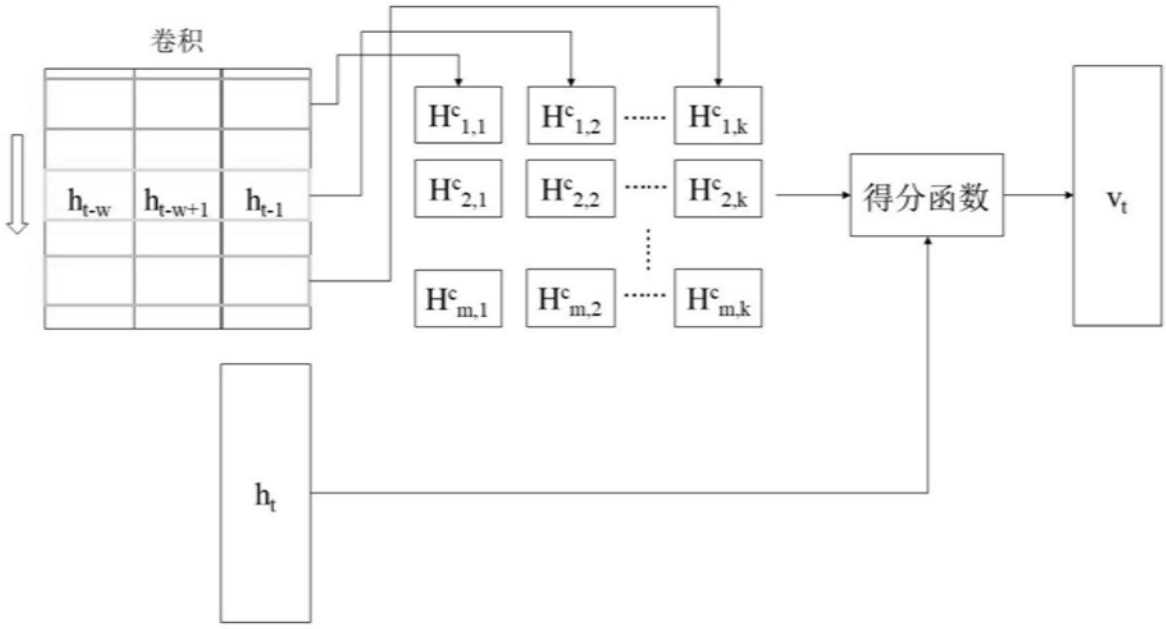

卷积
h_{t-w}
h_{t-w+1}
h_{t-1}
$H^c_{1,1}$
$H^c_{1,2}$
$H^c_{1,k}$
$H^c_{2,1}$
$H^c_{2,2}$
$H^c_{2,k}$
$H^c_{m,1}$
$H^c_{m,2}$
$H^c_{m,k}$
得分函数
v_t
h_t

图12

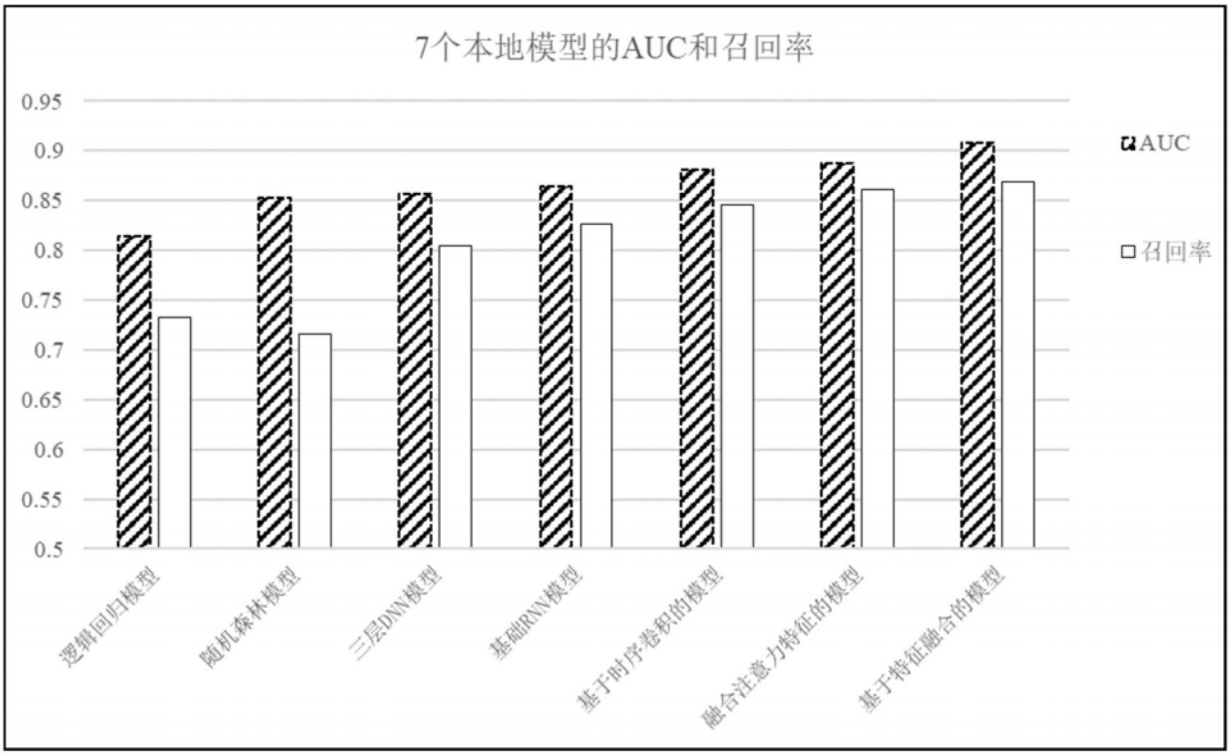

7个本地模型的AUC和召回率
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
AUC
召回率
逻辑回归模型
随机森林模型
三层DNN模型
基础RNN模型
基于时序卷积的模型
融合注意力特征的模型
基于特征融合的模型

图13

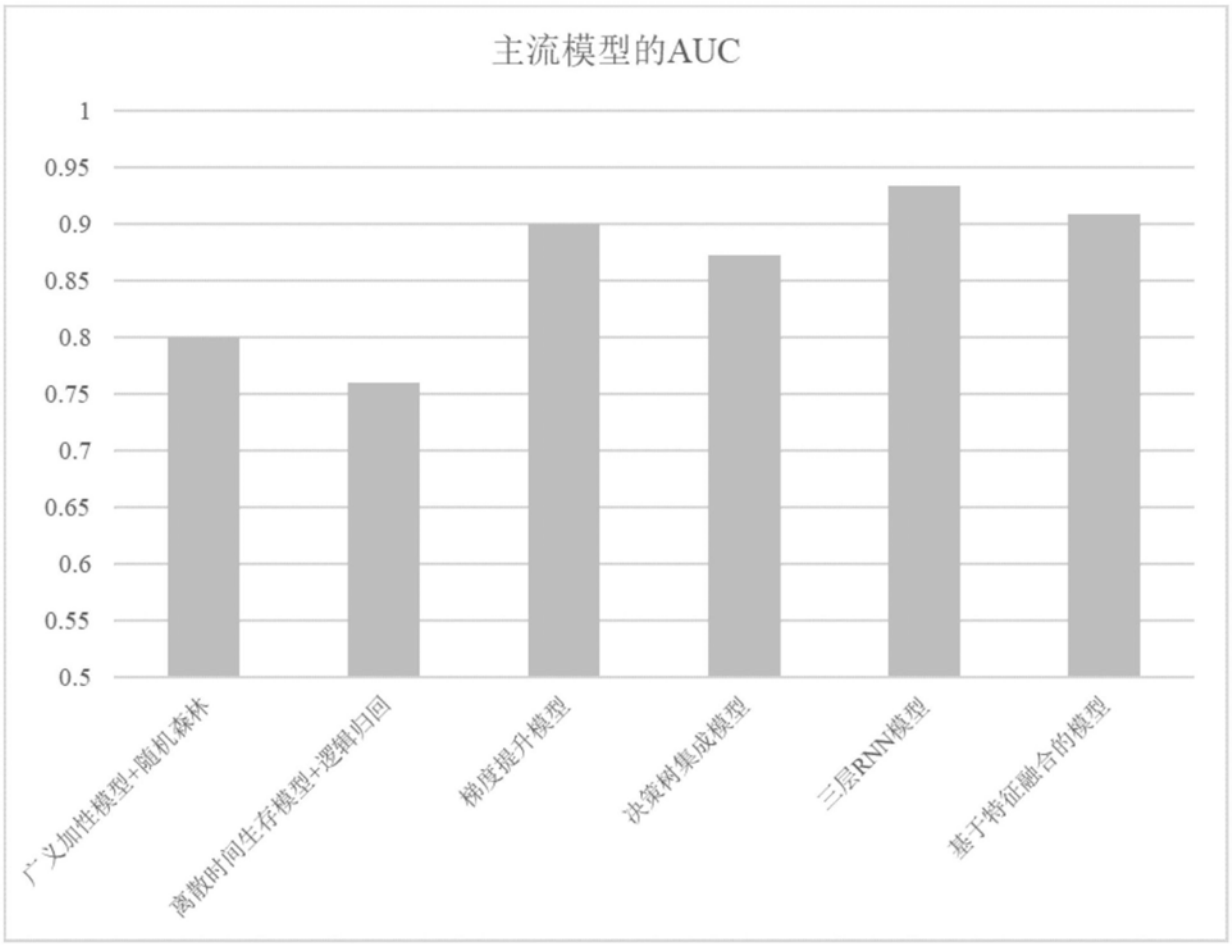

主流模型的AUC
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
广义加性模型+随机森林
离散时间生存模型+逻辑归回
梯度提升模型
决策树集成模型
三层RNN模型
基于特征融合的模型

图14