(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 110569511 A
(43)申请公布日 2019.12.13

(21)申请号 201910896154.6

(22)申请日 2019.09.22

(71)申请人 河南工业大学
地址 450001 河南省郑州市高新技术产业开发区莲花街100号河南工业大学科技处

(72)发明人 姜明伟 吴小雪 张庆辉

(51)Int.Cl.
G06F 17/27(2006.01)
G06N 3/04(2006.01)

权利要求书1页 说明书3页 附图1页

(54)发明名称

基于混合神经网络的电子病历特征提取方法

(57)摘要

基于混合神经网络的电子病历文本特征提取方法,依次包括如下步骤:(1)获取数据集;(2)数据预处理;(3)获取单词的词向量表示;(4)构建TextCNN模型获取邻近词汇间的关联,捕捉文本局部特征;(5)构建Bi-LSTM模型对获取的语义信息进行记忆,捕捉上下文关联信息,最大限度理解词汇的语义信息;(6)设计全连接层进行特征汇聚。本发明所述的方法提高了用于处理电子病历文本,获取电子病历文本特征的方法,从而获取文本特征语义信息。

CN 110569511 A

1.基于混合神经网络的电子病历特征提取方法,其特征在于:依次包括如下步骤:

(1)获取数据集:选取CCKS2017提供的电子病历文本数据集;

(2)数据预处理:分别对数据集进行文本分词、去停用词、词频统计、特征表示等手段,获得离散化的电子病历数据;

(3)获取单词的词向量表示:使用谷歌word2vec模型,将词从高维空间分布式地映射到低维空间且保留词向量之间的位置关系,从而解决向量稀疏和语义联系两个问题;

(4)构建TextCNN模型获取邻近词汇间的关联,捕捉文本局部特征;TextCNN模型包括卷积层和最大池化层;卷积层用于从不同大小的输入矩阵学习上下文相关的不同特征;参考局部感受野的思想,每个隐藏层节点只连接到某个足够小局部的输入点上,而不是全连接到每个输入点上,同时同一层中某些神经元之间的连接权重是共享的,从而大大减少需要训练的权值参数;卷积层使用多个$n \times h$的卷积核,与Word2vec模型的输出结果进行卷积操作,通过使用不同窗口大小的卷积核可以让网络自动提取出句子的不同特征;再将每一个卷积核与句子卷积得到的结果连接起来,得到卷积层的输出;池化层用于对特征进行下采样,增强模型的鲁棒性,并有效地提高模型的性能;在每次卷积过后,通过池化过程减小特征图尺寸,简化从卷积层输出的信息;本项目采用最大池化方法,对卷积层输出的每个向量取最大值,提取出最重要的特征信息,再连接成一个向量,得到池化层的输出;最大池化的方法能使网络自动提取句子中最有用的特征

(5)构建Bi-LSTM模型对获取的语义信息进行记忆,捕捉上下文关联信息,最大限度理解词汇的语义信息;构建的Bi-LSTM模型包括双向LSTM层、聚合层和最大池化层;双向LSTM层相当于特征抽取部分,通过构造两个LSTM神经网络来实现从两个相反的方向获取信息,更有利于从整体上捕捉句子的长依赖关系以及文本的深层语义表达,两个神经网络的输入一致;LSTM的优势在于具有三个特殊的门函数:输入门、遗忘门、输出门,通过这三种门来控制神经网络的记忆;聚合层将双向LSTM层得到的前向传播输出向量以及反向传播输出向量拼接起来;同时由于每条输入文本包含的词语数量不一致,通过池化操作也可得到定长的特征向量;

(6)设计全连接层进行特征汇聚;设计两个全连接层对TextCNN模型和Bi-LSTM模型提取的特征进行汇聚,生成最终用于慢性病分类的深度词向量特征;本项目在第一个全连接层前,使用Tensorflow框架中的concat( )方法对TextCNN和Bi-LSTM输出的特征进行融合;将融合后的特征作为第一个全连接层的输入,在第一个全连接层与第二个全连接层之间引入Dropout机制,每次迭代放弃部分训练好的参数,使权值更新不再依赖部分固有特征,防止过拟合。

基于混合神经网络的电子病历特征提取方法

技术领域

[0001] 本发明属于深度学习自然语言处理技术领域。

背景技术

[0002] 电子病历中包含了大量的数字和文本信息,是医务人员为患者开展相关治疗的实录。通过对信息进行抽取,得到有用的医疗数据,既可以为医疗提供决策支持,又可以为患者提供个性化诊断方案,实现精准医疗。深度学习模型因为可以自动从数据中提取特征成为了特征表示的研究热点,卷积神经网络(Convolutional Neural Network,CNN)和循环神经网络(Recurrent Neural Network,RNN)已经被证明是有效的语义组成模型。Zachary C.Lipton等首次提出评估使用长短期记忆(Long Short Term Memory,LSTM)网络识别多变量序列的临床病历的能力,其模型效果优于此前使用多层感知器的研究方法。R Miotto等提出了一种新的无监督深度特征学习方法,此方法可以在电子病历中获取一个病人的病理特征,使得针对性的临床预测建模更加方便。他们训练三层堆叠的降噪自动解码器辨别70万患者电子病历中的层次规律和依存关系。他们将得到的模型称为“深度患者”,其在严重糖尿病、精神分裂症及各种癌症的预测上表现出色。Nguyen P等提出了一种端到端深度学习系统Deepr,此系统可以从病历记录中提取病理特征并自动预测。其构建的“深度记录”可以提高临床诊断的准确性。WU Y等构建了一种针对中文电子病历命名体识别的深度神经网络。通过无监督学习将未标记的语料库生成词作为输入层,实验结果表明其模型优于其他CRF模型。吴嘉伟提出一种针对英文电子病历的实体关系抽取的特征学习方法,针对电子病历中文本结构稀疏的特点,将有限的上下文特征进行抽象表示,进而发掘出词与词之间的组合关系特征。Yang等人应用多层卷积神经网络对电子病历文本进行高层次语义理解,然后将其用于疾病诊断,取得了良好的效果。

[0003] 总的来说,现有的这些方法都是基于CNN或者RNN从病历记录中提取病理特征,但CNN模型侧重提取当前局部信息,RNN侧重保存句子的历史信息,它们在解释数据特征之间的相互影响作用时都存在无法结合时间及空间进行特征表示的缺点,而CNN和RNN的融合模型则可以使它们优势互补。

发明内容

[0004] 本发明旨在提供一种用于提取电子病历文本特征的混合神经网络方法。模型可以自动从电子病历文本中提取特征,通过词向量表示单词并通过神经网络来学习可变长句子的向量表示,从而可以很好地捕获句子的语义信息。CNN模型侧重提取当前局部信息,而RNN侧重保存句子的历史信息。它们的融合模型便兼顾了局部信息和上下文历史信息。但是在实际训练过程中RNN受限于梯度爆炸的影响,自RNN衍生的LSTM很好地解决了这个问题,采用TextCNN和Bi-LSTM的融合模型来提取电子病历文本中的特征信息。

[0005] 拟选取CCKS2017提供的电子病历文本,进行预处理。通过文本分词、去停用词、词频统计、特征表示等手段,获得离散化的电子病历数据。通过Word2vec的CBOW模型对电子病

历文本的语料进行训练,从而获得文本词向量表示。构建TextCNN模型获取邻近词汇间的关联,捕捉文本局部特征。构建Bi-LSTM模型对获取的语义信息进行记忆,捕捉上下文关联信息,最大限度理解词汇的语义信息。最后设计两个全连接层对TextCNN模型和Bi-LSTM模型提取的特征进行汇聚,生成最终用于慢性病分类的深度词向量特征。

附图说明

[0006] 图1为特征提取流程图。图2为“Bi-LSTM”结构图。图3为“TestCNN”模型结构图。

具体实施方式

[0007] 为了验证所提模型的有效性,在本文所选语料库上实验。具体包括以下步骤:

第一,对电子病历文本进行预处理。通过文本分词、去停用词、词频统计和特征表示,获得离散化的电子病历数据。基于jieba分词工具对每一个文本进行分词;基于百度提供的停用词表对其进行去停用词处理,继而进行去噪声处理。在去除噪声时,针对其中涉及到的特定术语缩写、URL、标点符号等字符串进行处理。把预处理后的电子病历数据集划分为训练样本和测试数据两部分,随机抽取数据样本的2/3进行模型训练,剩余1/3数据用来对模型的有效性进行评估。

[0008] 第二,构建相应的Word2vec模型进行词向量表示。谷歌开源工具Word2vec是一种快速有效地训练词向量模型的方法,分为CBOW和Skip-gram两种方式,采用CBOW模型对语料进行训练从而获得文本词向量表示。CBOW模型的输入是某一个特征词的上下文相关的词对应的词向量,输出是这个特征词的词向量。在找到每个词语对应的词向量之后,将每一个词向量堆叠形成词向量特征矩阵。

[0009] 第三,构建TextCNN模型获取邻近词汇间的关联,捕捉文本局部特征。在TextCNN卷积层分别使用3X100,4X100,5X100的卷积核各128个,步长stride大小为1,与Word2vec模型的输出结果进行卷积操作,通过使用不同窗口大小的卷积核可以让网络自动提取出句子的不同特征。并采用最大池化操作,使用三个池化层提取卷积后特征中的关键特征,剔除冗余操作,将多个池化操作的特征进行拼接生成一个固定维度的特征向量。

[0010] 第四,构建Bi-LSTM模型,将word2vec生成的词向量作为模型的输入,设置词向量维度为128维,分别输入模型的前向与后向。对获取的语义信息进行记忆,捕捉上下文关联信息,最大限度理解词汇的语义信息。最后将两个方向的输出结果拼接,最终生成模型输出。

[0011] 第五,在固定其他参数的情况下,分别比较128维,256维词向量维度,并分别取卷积网络的滑动窗口大小为3、5、7三种对比,同时dropout的比例分别对比0.2,0.4,0.5,0.7四种。结果显示当词向量维度为128维,滑动窗口大小为3、5,使用最大池化且dropout值为0.5时效果最好。使用同样的方法可知,当词向量维度为128维,模型的隐藏层大小为128时模型的特征提取准确度最高。

[0012] 第六,为了验证所提混合神经网络模型在电子病历特征提取上的效果,在同等条件下分别使用单TextCNN模型与单BiLSTM模型作为对比实验。将提取特征使用softmax做分类处理,获得最终分类准确率如表1所示。

模型	准确率/%

TextCNN模型	90.10
BiLSTM模型	92.15
TextCNN-BiLSTM模型	94.36

[0013] 由表中可以看出使用混合神经网络模型在同等条件下可以将提取效果提升2个百分点。以上结果表明是用混合神经网络模型结构在提取电子病历文本特征时能在一定程度上提升实验效果。

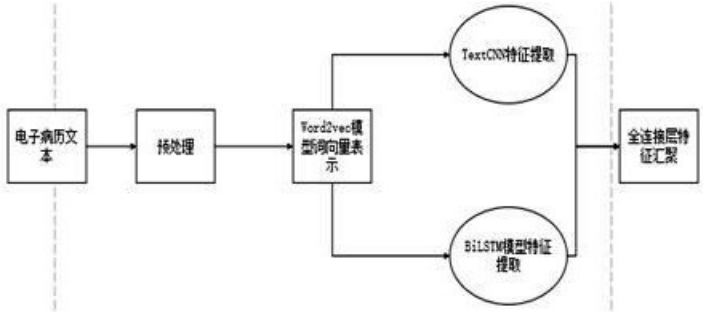
电子病历文本
预处理
Word2vec模型词向量表示
TextCNN特征提取
BiLSTM模型特征提取
全连接层特征汇聚

图 1

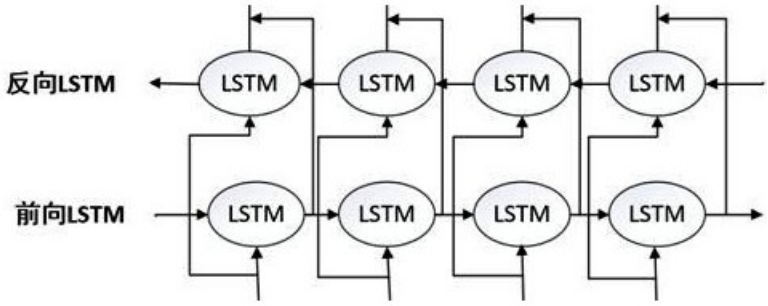
反向LSTM
前向LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM

图 2

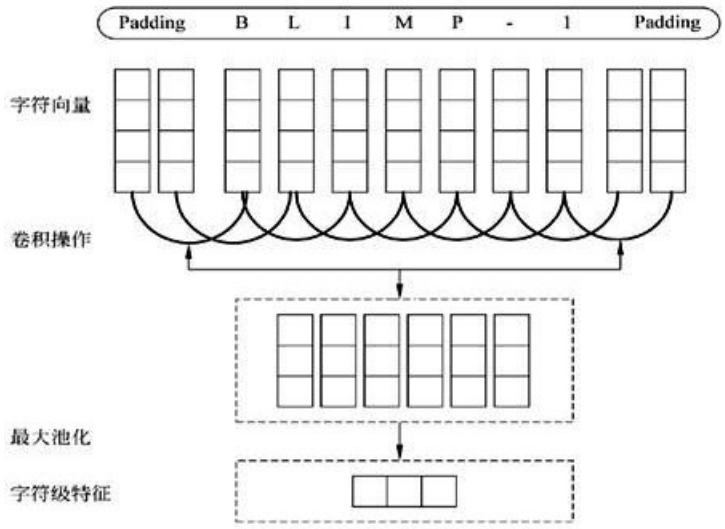
Padding
B
L
I
M
P
-
1
Padding
字符向量
卷积操作
最大池化
字符级特征

图 3