文章编号：2095-2163(2019)05-0154-04 中图分类号：TP393.08 文献标志码：A

面向呼吸内科智能诊断模型研究

胡金鹏，关 毅
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：随着科技的发展，智能医疗已经成为当下学界的热点研究内容。本文主要研究的是呼吸内科疾病的智能诊断，使用电子病历中的症状实体和异常检查结果实体来诊断患者可能患有的疾病。本文比较了不同的模型在该任务上表现，包括传统机器学习和深度学习。并且在深度模型中加入了不同的图表示学习方法以及提出了注意力机制来加强疾病和症状之间的联系。在实验中，本文提出的结合注意力机制和卷积神经网络以及外部向量获得了最优秀的表现。

关键词：深度学习；电子病历；实体识别；医疗信息；智能诊断

Research on intelligent diagnosis model for respiratory medicine

HU Jinpeng, GUAN Yi
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the development of science and technology, intelligemt medical treatment has become a hot research topic in the current academic circles. This paper focuses on the intelligent diagnosis of respiratory diseases, using symptomatic entities and abnormal test results entities in electronic medical records to diagnose diseases that patients may have. The paper compares the performance of different models on this task, including traditional machine learning and deep learning. In addition, graph representation learning methods are added to the deep learning and attention, which is used to strengthen the relationship between disease and symptoms. In the experiment, the model which combines attention, Convolutional Neural Network(CNN) and external vector achieve the best performance.

【Key words】 deep learning; electronic medical record; entity recognition; medical information; intelligent diagnosis

0 引 言

针对每年国内到各类医疗机构就医人群的绝对数量十分庞大的现状，医疗人员通常都会面临巨大的工作压力。而且绝大多数的就诊患者都分布在基层医疗机构。然而国内医疗资源的分配却存在着不均衡性[1]。近年来，随着科学技术的迅猛发展，政府对于智能诊疗技术也给予了高度重视与支持。国务院发布的《新一代人工智能发展规划》中，明确指出了智能诊疗技术的方向和前景，其中包括了未来在该方面的各种新模式和新手段，能够通过人工智能技术在医疗领域的广泛应用来建立先进的智慧医疗系统。例如，在手术方面可以通过智能机器人来代替医生，也可以通过一些智能的穿戴设备来随时监测病人的体征以及其它方面的信息，还可以用计算机实现影像识别，协助医生进行决策。时下，智慧医疗正逐渐成为热词，一方面是因为人工智能技术在近年间取得了可观的进步，另一方面来自日趋迫切的医疗需求，所以需要寻求合理医疗方案，以及建设有效智能诊断系统，来协助医生做出诊断，进而降低管理成本和提高医疗水平。这对于完善医疗保健系统和降低人口老龄化的压力都有着至关重要的现实意义。

在医疗诊断中，决策支持系统可以帮助医疗从业人员评估疾病风险。迄至目前，在诊断方面，各类研究成果也已相继涌现。Curiac 等人[2]使用贝叶斯模型去诊断精神类疾病。Lakho 等人[3]使用贝叶斯网络构建肝炎诊断决策支持系统，从知识模型中推断出结论，计算乙型肝炎、丙型和丁型肝炎疾病发生的概率。Kukreja[4]比较了神经网络、基于 C4.5 算法的贝叶斯网络以及反向传播等方法在哮喘诊断上的效果。Lin[5]使用分类回归树(CART)和案例推理技术(CBR)来构建诊断模型。Liang 等人[6]提出使用深度学习抽取电子病历中的特征来辅助医疗决策。Ogunleye 等人[7]将随机森林和局部回归相结合来增强节点输出的分辨率，在自闭症诊断中有着较为出色的表现。

本文提出的诊断模型是基于电子病历。研究

作者简介：胡金鹏(1996-)，男，硕士研究生，主要研究方向：自然语言处理、健康信息学；关 毅(1970-)，男，博士，教授，博士生导师，主要研究方向：人工智能、自然语言处理。

收稿日期：2019-06-12

中，抽取电子病历中的实体，包括症状、异常检查结果、疾病等，再通过症状和检查结果来推断出病者可能患有的疾病。为此，本文的主要研究工作可简述如下。

（1）实体识别。需要从自由文本的电子病历中抽取相应的实体。文中使用了 LSTM-CRF 模型，并且提出了将词向量和字符相结合的方法。

（2）电子病历中实体向量的生成。电子病历中的实体之间是存在关系，为此文中采用了图表示学习方法，同时采用了 deepwalk[8] 学习实体的向量表示。

（3）诊断模型的研究。本文对比了不同的模型在诊断上的表现，包括深度学习和传统的机器学习，而且提出了将 attention 引入到深度学习模型中的方法。

1　实体识别

实体识别是自然语言处理的信息抽取研究中的一个基础性的项目课题。总地来说，就是指在文本中抽取具有特定含义的信息，在 CoNLL－2002、CoNLL-2003 两届会议上将命名实体定义为包含特殊含义的短语，具体就是诸如人名、地名、机构名、时间等短语。本文中，实体识别研究主要是抽取电子病历中的相关实体。为此就会用到疾病、症状和异常检查结果等特征表述[9]。不同实体在电子病历中出现的实例详见表 1。

表 1　不同类型实体的实例

Tab. 1　Examples of different types of entities

实体	实例
疾病	左腕部骨折予以支具固定
	经检查诊断为关节炎
	院外就诊考虑“慢性支气管炎”经服药或输液治疗后可缓解
症状	患者无明显诱因出现多饮、多尿
	表现为头痛、头晕、眼花
	受凉后上述症状再发，呈阵发性咳嗽，咳少许淡黄色脓痰
	伴畏寒、发热、夜间阵发性呼吸困难
	双肺呼吸音粗，未闻及干细湿性啰音
	患者无明显诱因出现活动后呼吸困难
检查结果	少许癌组织出现在支气管
	胸片提示：双肺纹理增多
	患者多次查血常规示：血红蛋白低
	胸部 CT 提示：双侧胸膜稍增厚
	心电图示：窦性心动过速

实体识别是典型的序列标注任务。这里，采用的是基于双向 LSTM-CRF 的模型构建。LSTM 是改进的 RNN 单元，主要通过输入门、输出门和遗忘门来控制信息的传递。研究中采用模型的主题设计结构如图 1 所示。

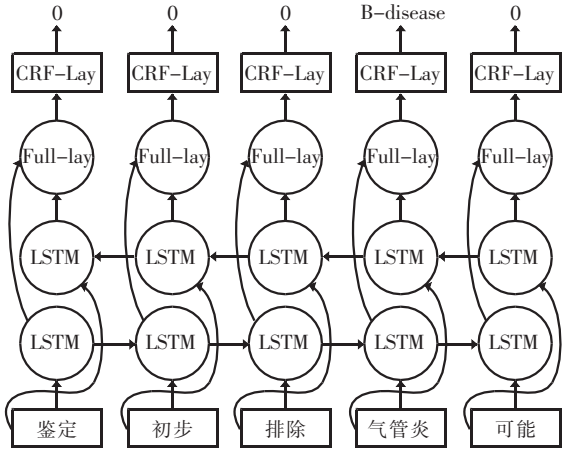

图 1　实体识别网络结构图

Fig. 1　Network structure of entity recognition

同时，还在 Embedding 层做出了改进，将基于字符的向量加入到每个词的表示中。基于字符的向量生成如图 2 所示。

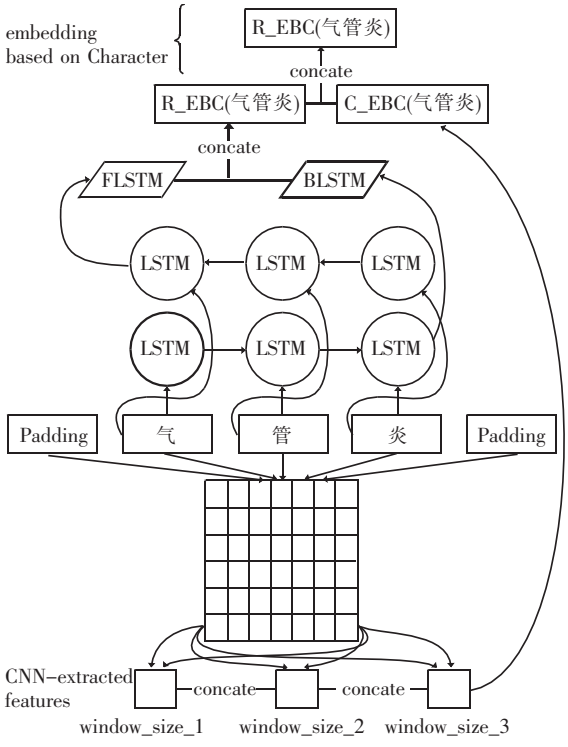

图 2　基于字的向量的生成

Fig. 2　Character-based embedding generation

由图 2 分析得知，该向量由 2 部分组成。一部分由 LSTM 产生，另一部分由 CNN 产生。每个词都是由字组成的序列，故而可使用 LSTM 来抽取词的序列特征。考虑到 CNN 在抽取局部特征有着较强

的能力[10],本次研发中使用了 CNN 来抽取每个字的 n-gram 特征。

2 基于深度学习的诊断模型

基于深度学习诊断模型结构如图3所示。在图3中,$P = [w_1, w_2, \cdots, w_n]$ 为一个患者的所有症状的索引。研究中,需要通过 Embedding 层将这些症状转化为向量。在将这些向量送入卷积层之前,需要对这些向量进行 Attention 处理。在此,使用的向量是疾病向量和症状之间加入了注意力机制。通过将不同疾病生成一个疾病向量表,在训练过程中可将每个训练数据的标签从疾病的向量矩阵中根据索引值获得相关的向量,接着将症状向量和疾病向量加以 Attention 处理。对该过程可阐释详述如下。

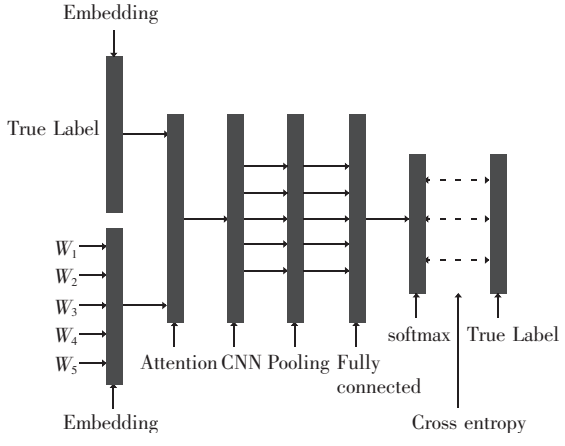

图3 基于深度学习诊断模型结构

Fig. 3 Diagnostic model structure based on deep learning

假设该病例的目标标签为 d,d 为疾病向量,也就是标签对应的向量。P 中每个症状都与疾病向量 d 做内积,再经过 softmax 层获取 α 值,每一个症状都会有与之相对应的权重。其数学公式可写作如下形式:

$$\alpha = \text{softmax}(P \cdot d), \tag{1}$$

将原来的症状向量分别乘以 α 中相对应的数值,获得经过 Attention 处理后的症状向量 P'。其数学公式可表示为:

$$P' = P * \alpha \tag{2}$$

本次研究中使用的 Attention 的设计流程详见图4。在此基础上,将 P' 传入到 CNN 中执行后续的特征提取过程。通过 CNN 抽取不同症状的 n-gram 特征,在池化层中采用了最大池化法。究其原因则在于:一般情况下,数值最大的特征表示的就是最重要的特征。或者是直接将 P' 输入到全连接层中来进一步实现提取特征。

在前述研究结束后,最后一层就是 softmax 层。文中选用的是交叉熵损失函数。这是因为交叉熵函数的训练速度相对于平方差损失函数来说,在训练速度上会更快。

与此同时,在训练模型中,本文中选择的是反向传播算法,通过反向传播不断改进网络中的参数,使其能够更好地拟合训练数据。

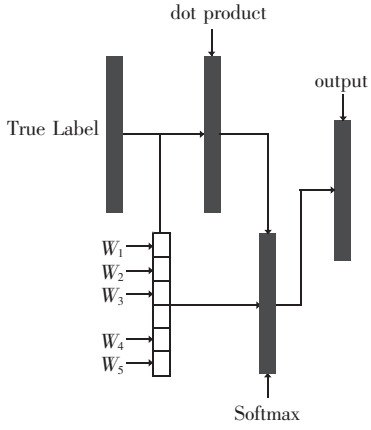

图4 注意力机制

Fig. 4 Attention mechanism

3 实验结果及分析

研究中,对于不同实体识别方法的效果对比见表2。由表2可以看到,当加入字符级的词向量时,模型的表现为最佳,达到了0.959 7。而基于 LSTM 的实体识别模型的表现最差,则是因为 LSTM 没有考虑到输出标签之间的联系,因而会出现一些不可能的错误。例如,"B-diseasae"后面的词的标签不可能是"I-test"。"B-diseasae"后面的标签只能是"I-disease"、"B-XX"或者"O"。但通过仿真结果讨论后可知,如果将 LSTM 后面加上 CRF 层就会避免该类错误,还可以发现字符级特征在实体识别上也是有效特征,能够更好地表征各个单词的含义,从而提高了实体识别的效果。

表2 实体识别的实验结果

Tab. 2 Performance of different entity recognition models

方法	*F*1
LSTM	0.934 5
CRF	0.944 5
LSTM+CRF	0.951 4
LSTM+CRF(char)	**0.959 7**

不同诊断模型的实验结果对比见表3。由表3可以看到,在 top1 的表格中,CNN-attention-deepwalk 获得了最好的效果,这说明 CNN 能够有效抽取症状特征,当加入外部词向量时也大大提高了

模型的准确度。但就总体来说，所有算法的运行效果都较差，这是因为大多数疾病的数据都很稀疏。本次研究得到的疾病频率统计结果如图 5 所示。由图 5 可知，高达 79% 的疾病在本文选取的数据中出现的次数都不超过 5，频数超过 30 的疾病仅占据所有疾病的 8%。

表 3　不同诊断模型的实验结果

Tab. 3　Performance of different disease diagnosis models

方法	Top1	Top2	Top3
RandomForest	0.558 3	0.730 0	0.800 0
DecisionTree	0.538 3	0.693 3	0.776 6
KNeighbors	0.493 0	0.635 0	0.723 0
CNN	0.578 3	0.720 0	0.790 0
DNN	0.540 0	0.701 7	0.783 3
CNN+Attention+deepwalk	0.595 0	0.748 3	0.813 3
DNN+Attention+deepwalk	0.575 0	0.715 0	0.801 7

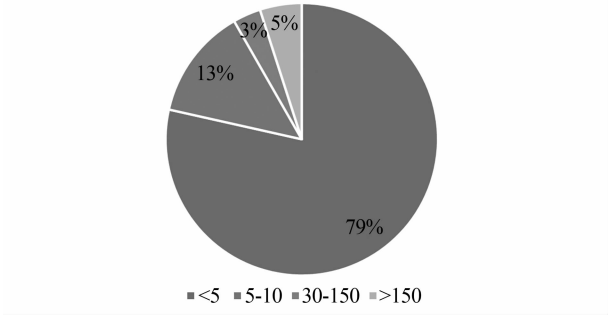

图 5　疾病频率统计图

Fig. 5　Disease frequency chart

研究过程中发现，当从 top1 指标转换为 top2 和 top3 时，模型的准确率有了显著提升，这就说明本文研发的模型能够有效地提取特征并做出准确诊断。

4　结束语

随着智慧医疗热潮的到来，越来越多的人开始重视人工智能在医疗领域的应用。每年的就诊人次也在快速地增长，当前的医疗资源建设、尤其是基层医院，已难以满足人民群众的就医诊治需求。智能诊断发挥着越来越重要的作用。本文提出了使用人工智能技术对呼吸内科疾病进行诊断，并对比了不同模型在该任务的表现。同时，也提出了使用 CNN 以及 Attention 机制来诊断呼吸内科疾病，在所有模型中获得了最优的表现。但是整体的准确度还未能达到实际应用的水平。此后，还需要采集更多的数据来训练模型，也要解决数据倾斜的问题。

参考文献

[1] 梁玮佳，唐元懋. 我国卫生资源配置的空间非均衡研究[J]. 卫生经济研究，2018(9)：66-71.

[2] CURIAC D I, VASILE G, BANIAS O, et al. Bayesian network model for diagnosis of psychiatric diseases[C]// Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces. Croatia：IEEE, 2009：61-66.

[3] LAKHO S, JALBANI A H, VIGHIO M S, et al. Decision support system for hepatitis disease diagnosis using bayesian network[J]. Sukkur IBA Journal of Computing and Mathematical Sciences，2017, 1(2)：11-19.

[4] KUKREJA S. A comprehensive study on the applications of machine learning for the medical diagnosis and prognosis of Asthma[J]. arXiv preprint arXiv：1804.04612v1，2018.

[5] LIN R H. An intelligent model for liver disease diagnosis[J]. Artificial Intelligence in Medicine，2009, 47(1)：53-62.

[6] LIANG Z, ZHANG Gang, HUANG Xiangji, et al. Deep learning for healthcare decision making with EMRs[C]//2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Belfast, UK：IEEE, 2014：556-559.

[7] OGUNLEYE A, WANG Qingguo, MARWALA T. Integrated learning via randomized forests and localized regression with application to medical diagnosis[J]. IEEE Access, 2019, 7：18727-18733.

[8] PEROZZI B, Al-RFOU R, SKIENA S. Deepwalk：Online learning of social representations[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM, 2014：701-710.

[9] 杨锦锋，关毅，何彬，等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报，2016，27(11)：2725-2746.

[10] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv：1408.5882, 2014.

(上接第 153 页)

[2] 毛承胜. 基于贝叶斯决策理论的局部分类方法研究及其应用[D]. 兰州：兰州大学，2016.

[3] LI Xiaowei, ZHAO Qinglin, HU Bin, et al. Improve affective learning with EEG approach[J]. Computing and Informatics, 2012, 29(4)：557-570.

[4] LU Yifei, ZHENG Weilong, LI Binbin, et al. Combining eye movements and EEG to enhance emotion recognition[C]//IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina：AAAI，2015：1170-1176.

[5] 张春燕. 基于自然语言处理的文本分类分析与研究[D]. 赣州：江西理工大学，2011.

[6] GRAUMAN K, DARRELL T. The pyramid match kernel：Discriminative classification with sets of image features[C]// IEEE International Conference on Computer Vision. Beijing, China：IEEE, 2005：1458-1465.

[7] CAO Jun, AHMADI M, SHRIDHAR M. Recognition of handwritten numerals with multiple feature and multistage classifier[J]. Pattern Recognition, 1995, 28(2)：153-160.

[8] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[9] DILRUKSHI I, ZOYSA K D. Twitter news classification：Theoretical and practical comparison of SVM against Naive Bayes algorithms[C]//International Conference on Advances in ICT for Emerging Regions (ICTer). Colombo Sri Lanka：IEEE, 2013：278.