面向疾病相关关联抽取的深度语义特征研究

康旭琴[1],吴 偶[2],王 磊[1],张 音[1],杨 帅[1]
KANG Xuqin[1], WU Ou[2], WANG Lei[1], ZHANG Yin[1], YANG Shuai[1]

1.军事医学科学院 卫生勤务与医学情报研究所,北京 100850
2.天津大学 应用数学中心,天津 300072
1.Institute of Health Service and Medical Information, Academy of Military Medical Sciences, Beijing 100850, China
2.Center for Applied Mathematics & School of Mathematics, Tianjin University, Tianjin 300072, China

KANG Xuqin, WU Ou, WANG Lei, et al. Research on deep semantic features for disease-association relation extraction. Computer Engineering and Applications, 2018, 54(8):260-264.

Abstract: It is of great reference value for the research of prevention and cure of disease to find out the beneficial or harmful factors affecting the disease from a large number of biomedical literatures. However, it is difficult to identify the bottlenecks that are difficult to continue to improve when the accuracy of the classification is improved to a certain level by using the traditional machine learning method. In order to improve the performance of the classification task in biomedical field, the hybrid method of convolution neural network method and Support Vector Machine(SVM) is used by data with the two factors which are beneficial and harmful to disease. Ultimately, the method achieves better performance than the traditional machine learning, with the accuracy of classification from SVM increasing from 90.44% to 94.38%, so as to better identify the factors that affect the disease.

Key words: relation extraction; classification problem; deep learning; machine learning; convolutional neural network; Support Vector Machine(SVM)

摘 要:从大量生物医学文献中找出影响疾病的有利因素和有害因素对于疾病的防治研究方向有着重要参考意义。然而,识别疾病影响因素的二分类问题在用传统的机器学习方法进行分类时正确率提升到一定水平后遇到瓶颈难以继续提高。为了提高生物医学领域二分类问题模型的分类性能,利用对于疾病有利和有害的两种因素,采用基于卷积神经网络与支持向量机(SVM)相结合的方法,最终达到超过传统机器学习的性能,使分类的准确率从SVM最佳的90.44%提升到94.38%,从而更好地识别疾病的影响因素。

关键词:关联抽取;分类问题;深度学习;机器学习;卷积神经网络;支持向量机
文献标志码:A **中图分类号:**TP391.1 doi:10.3778/j.issn.1002-8331.1709-0336

1 前言

医学的发展经历了长久以来倚靠专家的经验医学、参考类似病例治疗的循证医学,开始走向个体化针对性治疗的精准医学[1]。除了医生的个人经验,从文献中发现疾病的影响因素也是现代医学发展的重要方面。近年来,生物医学领域的权威数据库PubMed已收录2 700余万条文献记录[2],导致科研人员和医生难以从海量生物医学文献中发现高质量、可用性的知识。所以,如何快速、高效地从文献中发现疾病的影响因素成为医学的重要发展方向。近年来受到广泛关注的可用于文本挖掘及知识发现的深度学习方法有着强大的计算表达能力,在特定性能评测中已超越了传统机器学习方法。因此,面对海量生物医学文献的挑战,要充分探索利用新技术新方法发现潜在的知识,挖掘疾病相关的因素并加

基金项目:国家重点研发计划课题(No.2016YFC09019002);北京市科技计划课题(No.Z171100003217038)。
作者简介:康旭琴(1991—),女,博士研究生,主要研究方向为生物医学文本挖掘、生物医药科技情报,E-mail:kangxq@foxmail.com;吴偶(1982—),男,博士,教授,主要研究方向为数据挖掘与机器学习;王磊(1971—),通讯作者,女,博士,研究员,主要研究方向为文本资源获取、医学信息分析及深度情报挖掘等;张音(1980—),通讯作者,女,博士,副研究员,主要研究方向为生物医药科技情报研究;杨帅(1992—),硕士研究生,主要研究方向为生物医药科技情报。
收稿日期:2017-09-25 **修回日期:**2017-11-20 **文章编号:**1002-8331(2018)08-0260-05

计算机工程与应用 www.ceaj.org

以利用,这对于精准医学发展具有重要的推动作用。融合多种文本特征、选取合适的方法和模型,实现高性能的关联抽取,并在公开数据中对疾病相关因素给出有意义的提示非常重要。对疾病相关的因素进行深度研究,从中提取有意义的相关关系,为科研假设提供思路和线索,给出借鉴和参考。

语义关联抽取,指从文本中自动识别两个命名实体之间的关联。从文本中抽取出的信息元素称为命名实体[3]。生物医学领域的关联抽取是要实现从生物医学文本中识别出生物医学命名实体(疾病、药物、基因、蛋白质等),提取实体之间的语义关联并形成关系网络。常见的实体语义关联抽取技术方法包括以下四种:基于词典驱动的关联抽取、基于模式匹配的关联抽取、基于本体的关联抽取以及基于机器学习的关联抽取。目前,机器学习方法成为关联抽取的主要方法。

消息理解评测会议(Message Understanding Conference,MUC)和后来的自动内容抽取评测会议(Automatic Content Extraction,ACE)、国际文本分析会议(Text Analysis Conference,TAC)促进了关联抽取技术的进步。关联抽取的评测会议主要包括JNLPABA(Joint Workshop on Natural Language Processing in Biomedicine and Its Applications和BioCreAtIve(Critical Assessment of Information Extraction Systems in Biology),应用最广泛的评测数据集包括GENIA语料库和GENETAG语料库。Qian等人[4]基于树核进行了蛋白质-蛋白质关联抽取。冯钦林等人[5]使用半监督学习的Tri-training的方法,进行疾病-病症和病症-治疗物质的关系抽取研究,利用大量未标注数据辅助少量有标注数据进行训练提高分类性能。杨晨浩[6]使用基于RNTN和RNN方法进行中文电子病历实体修饰与关系抽取研究,也得到有益的实验结果。Zeng等人[7]发表在COLLING 2014上的论文,使用卷积神经网和词向量,加入了位置特征,在不需要提取复杂特征的情况下,也取得了超过传统方法的分类性能。

2　方法

在有监督学习中,关联抽取问题可以抽象为分类问题,即判断二者的关系为预设选项中的哪一类[8]。本实验中,通过提取丰富特征输入到支持向量机中与直接提取独热编码[9]特征利用支持向量机分类进行比较。具体思路是首先使用深度学习方法,利用词向量工具训练出词向量,再将经过词向量转化的句子输入到卷积神经网络中,进行卷积与池化,然后进行全连接得到深度学习特征(深度学习进行特征提取的网络结构见图1);再将顶层特征作为整体输入到支持向量机模型中进行分类,最终得到的分类效果超过直接用支持向量机得到的分类结果。以下从每一环节进行展开阐述。

2.1　词向量

2.1.1　文本预处理

为保证后续关联抽取的质量,对原始的文本数据需要进行必要的文本预处理[10]。其中,符号的处理是重要的部分。为了保留符号的语义信息,本实验只把符号跟单词分隔开,并未直接删除。此外,所有的单词进行了小写化处理,避免将同一单词的大小写形式作为不同词处理。

2.1.2　词向量工具

词向量是指用一组向量来表示单词。主流的词向量工具已经可以较好地表达单词的语义信息,使计算机处理文本便捷实用[11]。本实验在文本预处理的基础上,将所有单词用相对于独热(one-hot)词向量表示方式更低维、稠密且带有语义的实数向量来表示。单个词的维度用 d^a 表示,所有数据中不重复的单词个数即词汇量总数为 V 。

2.2　文本表示

得到单个词的词向量后,考虑用其他更加丰富的信息来表达文本。实体附近的词往往对于确定实体关系有着重要的信息,因此当前词对于句子中两个实体的相对位置也纳入考虑[7]。例如,直观表示句子"It also seems that some pharmacologic properties of esomeprazole [e1] are actually better for the treatment of GERD [e2]."中,treatment相对于两个实体e1和e2的距离分别是6和–2(实际中,位置向量维度不一定是实际词数距离,而是打散的多维数值)。每个词的相对每个实体的位置向量维度用 d^b 表示。

有了单个词向量和每个词的位置信息之后,利用二者共同表示一个单词(见图1文本表示层),即每个词 w 的维度 $d=d^a+d^b\times 2$ 。含有 m 个单词的句子可以表示为 $\boldsymbol{s}=\{w_1,w_2,w_3,\cdots,w_m\}$ 。

2.3　卷积

为了获取局部基础特征,对句子的文本表示(矩阵)进行卷积运算[12]。选取卷积核的列数即卷积滑动窗口长度为 l 。这样就得到维度为 $d\times l$ 的单个卷积核。为了得到所有边缘信息,句子向量首尾都需要补充维度为 $d\times(l-1)$ 的补丁再进行边缘卷积,补丁向量此处统一设为全零向量。这样,滑动窗口的个数为 $m+l-1$ (即使用单个卷积核之后得到 $m+l-1$ 的行向量),第 i 个滑动窗口内的 w 个词序列就是 $q_i=w_{i-l+1:i}$,其中 $i\in[1,m+l-1]$ 。此外,为了提取句子的不同特征,选取卷积核的数量为 d^c 。如果用 W 表示所有卷积核,那么第 j 个卷积核卷积之后得到的特征图 $p_j=[Ws+\boldsymbol{b}]_j$,其中 $j\in[1,2,3,\cdots,d^c]$,$\boldsymbol{b}$ 是偏置向量,用于调整卷积后的向量。

2.4　分段池化

为了进一步提取显著特征,卷积操作之后要进行池

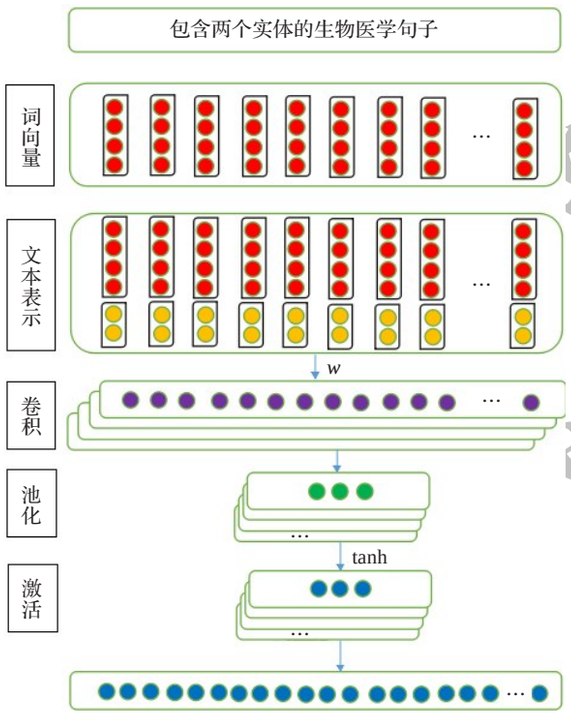

图1 深度学习进行特征提取的网络结构图

化。文本处理中不同句子长度卷积之后得到的特征图维度不同,为了方便后续统一维度的运算处理,通常采用最大池化方式。在最大池化的基础上,已有实验证明根据2个实体切分句子为3部分之后,在每一部分进行最大池化处理可以在保证统一维度的前提下取得更好的效果[8]。分段最大池化后的第 j 个特征图可以表示为 $x_j^k=\max(p_j^k)$,其中 $k\in[1,2,3]$。x_j 表示第 j 个特征图分段最大池化后向量的顺序连接。

2.5 非线性激活

在卷积和最大池化之后进行非线性激活,使模型具有更强的表达拟合能力。常用的激活函数有双曲正切函数tanh、sigmoid函数、Rectified Linear Units(ReLu)[13]。至此得到每个句子的特征维度为 $d^c\times 3$。为了验证这些丰富特征的作用,提取这些向量作为特征输入支持向量机(SVM),进行关联抽取。

2.6 支持向量机

SVM是由Vapnik等人提出的一种基于统计的学习方法[14],从线性可分情况下的最优分类面发展而来,它是对结构风险最小化归纳原则的近似。SVM中利用最优分类线来区分出不同样本,即不仅要正确分开不同类别,还要使分别通过距离分类线最近的多类样本的平行线之间的分类间隔最大。LIBSVM是快速有效的SVM模式识别与回归的软件包,成为本实验用SVM分类的工具[15]。

按照LIBSVM要求的格式,输入每一样本的标签,索引和对应的从深度卷积神经网络得到的特征值三部分数据进行模型训练,分别采用线性核与非线性核进行多次实验,得到最优参数,并在测试集上进行分类效果测评。

2.7 算法

输入:输入一个由 m 个单词组成,且标记出实体的句子,$s=\{w_1,w_2,w_3,\cdots,e1,\cdots,e2,\cdots,w_m\}$。

步骤1 在准备好的词向量表中查找每个词的对应向量,得到矩阵 $X_1\in d^a\times m$。

步骤2 给每个词的向量表示增加位置信息,得到矩阵 $X_2\in d\times s, d\in(d^a+d^b\times 2)$。

步骤3 用 d^c 个 $d\times l$ 的矩阵与 X_2 分别进行补丁卷积运算,得到矩阵 $X_3\in d^c\times(m+l-1)$。

步骤4 以句子中的两个实体划分句子为三部分,每一部分进行最大池化运算,得到矩阵 $X_3\in d^c\times 3$。

步骤5 对 X_3 进行非线性激活运算,得到矩阵 $X_4=\tanh(X_3)$。

步骤6 将 X_4 进行平铺,得到行向量 $k\in d^c\times 3$。

步骤7 将向量 k 作为特征,输入SVM进行关联抽取实验。

输出:得到每一个句子两个实体之间的关系(对疾病实体有益或有害)。

3 实验及评价

本实验数据源自PubMed,由Rahul Verma和Spiro Razis整理得到疾病相关的23 926条句子(以下简称本数据)[16]。每条句子中标记出疾病(实体1)和对其产生作用的另外一种实体(实体2),这两种实体的关系分为实体2有益于实体1和实体2有害于实体1两类。其中,有益的数据有12 293条(标记为正例),有害的数据11 633条(标记为负例),正负样例数据量均衡。以Rahul Verma和Spiro Razis实验中的最好方法核函数为rbf标准SVM为参照,数据划分为训练集84%(20 151条数据)、测试集16%(3 775条数据)。

3.1 评估指标

分类实验中常见的评价指标有准确率(precision,p)、召回率(recall,r)、正确率(accuracy,acc)、$F1$值($F1$ value)[17]。准确率为检索到的相关文档除以所有被检索到的文档得到的比率。召回率,也叫查全率,是检索出的相关文档数和文档库中所有的相关文档数的比率。$F1$值是二者的调和均值。本实验中正负样例均衡,用简单的精确度指标已经能够说明结果的主要性能,因此使用的主要评价指标是正确率(accuracy,acc)。在词向量预实验阶段用到$F1$值作为评价指标。二分类实验中样本类别分为P类(positive,阳性正例)和N类(negative,阴性负例),预测结果类别分为T(ture,真)和F(false,

假）。组合之后，把正例样本预测为正例的数量记为TP，把正例预测为负例的数量记为FN；把负例预测为正例的数量记为FP，把负例预测为负例的数量记为TN[18]。在此基础上，正确率的计算方式如下：

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN}$$

3.2　实验设置

按照第二章的方法开展本实验，实验条件设置环节涉及深度学习方法部分的词向量、文本表示、卷积、分段池化，支持向量机部分。

在词向量环节，分别选择调用已有词向量包和重新训练词向量。调用部分分别选择Yankai Lin等人所选通用领域的NEW YORK TIMES语料库在word2vec中生成的词向量[19]，Biomedical natural language processing[20]中预训练的分别基于PubMed Central，wikipedia-PubMed-and-PMC的词向量（大小分别为1.90 GB，4.11 GB）。重新训练词向量分别采用本数据的23 926条句子，PubMed中疾病为主题的原始题录和摘要，PubMed中疾病为主题的清洗过的题目摘要。训练词向量部分分别设置单个词维度为50、100、150、200、250、300、500，单词准入阈值为1、2、5，分别进行训练，并在原始数据2大类细分的6小类中进行预实验，从而根据结果筛选出本例中最适合的词向量，用于后续的二分类实验。

在文本表示环节，根据以往经验选择每个词相对实体位置的向量维度 $d^b=5$。在卷积环节，选择用Yankai Lin实验中得到的最优参数，设置滑动窗口数量 $l=3$，卷积核的数量设为 $d^c=230$。分段池化中一个句子由2个命名实体分为3段，分别进行最大池化。非线性激活环节采用双曲正切函数tanh，使每一个分量落在 $[-1,1]$ 之间。训练次数 N={200，300，400，500，600，700，800，900，1 000}，学习率 α={0.001，0.01，0.02，0.03，0.04，0.05，0.06，0.07，0.08，0.09，0.1，0.2}，多次训练之后取最佳结果。然后将每个样本690维的数据作为输入SVM的初始特征。除了将深度学习特征放入SVM模型中之外，也在原始SVM模型中反复训练调参（线性核、rbf核、多项式核、sigmoid核[15]），主要对当前性能稳定良好的rbf核部分进行了多种尝试，C={0.01，0.1，1，10，100}、*gamma*={0.01，0.1，1，10，100}，并在此基础上进行网格寻优 $C=(2^{-8},2^{8})$、$gamma=(2^{-8},2^{8})$，得到利用SVM方法时的最佳性能，并进行比较。

3.3　实验结果

随着训练次数的增加，准确率和召回率在此消彼长中逐渐趋于稳定平衡趋势（见图2分6类时不同词向量下的准确率/召回率）。实验结果如下：第一，用训练成熟的PubMed词向量（系列2和系列3）处理生物医学领域问题，明显优于不成熟训练的生物医学小词库训练得到的词向量（实验的6分类问题中 F 值最高约56.58%）。第二，系列2用基于PubMed的词向量和系列3用基于PubMed和Wiki的词向量结果没有统计学差别。第三，系列1与系列4相比，可以发现即使不成熟训练的生物医学领域词表性能也要优于通用领域训练成熟的词向量。第四，系列4只有50维，相比于200维的系列1、系列2、系列3能够训练得更快。

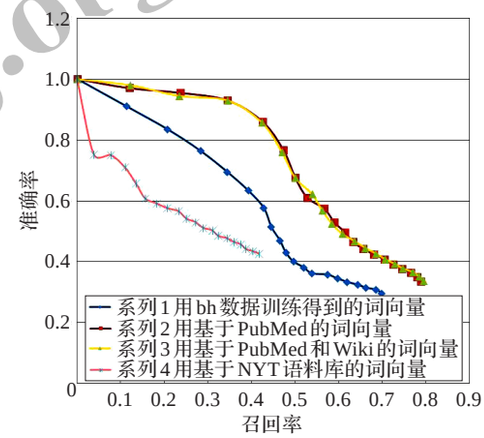

图2　分6类时不同词向量下的准确率/召回率

词向量预实验最终筛选出使用基于PubMed的词向量，在不损失性能的前提下用相对简单的词表（即图2中的系列2所用词表）。

训练次数 $N=1\,000$，学习率 $\alpha=0.07$ 时，预实验结果性能达到稳定的最佳。$N=500$ 时，结果开始趋于稳定。在将深度学习方法所提取的特征放到SVM的rbf中时，正确率 *acc* 达到94.38%（此时迭代次数为11 408次，3 775例测试集中正确的为3 563例），超过利用原始SVM线性核时应用one-hot编码简单特征77.22%的正确率、应用深度学习提取的丰富特征时90.65%的正确率以及SVM中采用rbf时最高90.44%的正确率（表1）。

表1　采用SVM线性和非线性核时不同特征下的准确率　%

模型	特征	
	one-hot编码简单特征	深度学习提取的丰富特征
SVM（linear）	77.22	90.65
SVM（rbf）	90.44	94.38

3.4　实验分析与讨论

词向量预实验部分的结果表明以下几点：第一，只要基于训练成熟的PubMed词向量，其他的加减语料几乎没有影响。第二，实际任务与词向量中语料的强相关性。

关联抽取实验可以说明以下几点：第一，训练次数 N 与学习率 α 不宜单独调整，单独调整时容易陷入局部最优，多种组合尝试才更容易取得最佳效果。第二，传统机器学习方法SVM中核函数（linear，poly，sigmoid，

计算机工程与应用 www.ceaj.org

rbf)的选择和参数设置(*C*, *gamma*)对疾病相关关联抽取结果的影响也很明显(表1)。第三,无论在线性核还是非线性核模型中,深度学习在特征提取环节对于提升性能作用明显(表1)。第四,即使在这样简单的疾病相关二分类问题中,深度学习方法在性能上仍有相对于传统SVM方法的优势(表1)。

但是,本文只讨论了对于疾病有益和有害这种典型的简单二分类问题的判别,在疾病的更多种分类问题尚未涉及;在某种特定疾病中的分类研究也未涉及;深度语义特征的选取也有局限。此外,入选的对于疾病有益或有害的样本基于特定文献库,正负样例相对平衡,但实际的文献库中可能正例报道会更多,也有一定的局限。

综上,本文在尝试了不同词向量对深度学习方法结果影响的基础上,将深度学习与传统机器学习方法结合运用。深度学习部分的特征通过自动学习得来,中间过程涵盖了变长卷积、分层最大池化等逐层提取特征的技术。结果表明这样的处理能够提升该疾病相关关联抽取的整体性能。后续将继续发现更多有益的特征,利用深度学习方法在具体病种中进行关联抽取实验以及开展多种分类的研究,进一步对比验证其在细分领域的效果。

参考文献:

[1] 赵晓宇,刁天喜,王磊,等.美国"精准医学计划"解读与思考[J].军事医学,2015,39(4):241-244.

[2] National Center for Biotechnology Information, U.S.National Library of Medicine.Home-PubMed-NCBI[EB/OL].[2017-07-02].https://www.ncbi.nlm.nih.gov/pubmed.

[3] 杨建明.关系抽取方法研究[J].电子技术,2009(4):36-41.

[4] Qian L,Zhou G.Tree kernel-based protein-protein interaction extraction from biomedical literature[J].Journal of Biomedical Informatics,2012(3).

[5] 冯钦林,杨志豪,林鸿飞.疾病-病症和病症-治疗物质的关系抽取研究[J].计算机工程与应用,2017,53(10):251-257.

[6] 杨晨浩.基于深度学习的中文电子病历实体修饰与关系抽取研究及算法平台开发[D].哈尔滨:哈尔滨工业大学,2016.

[7] Zeng D, Liu K, Lai S, et al.Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, The 25th International Conference on Computational Linguistics: Technical Papers, 2014: 2335-2344.

[8] 黄勋,游宏梁,于洋.关联抽取技术研究综述[J].现代图书情报技术,2013(11):30-39.

[9] 鲁亚平.面向深度网络的自编码器研究[D].江苏苏州:苏州大学,2016.

[10] Zhao Z, Yang Z, Lin H, et al.A protein-protein interaction extraction approach based on deep neural network[J]. International Journal of Data Mining & Bioinformatics, 2016,15(2):145-164.

[11] Rong X.word2vec parameter learning explained[J].Computer Science,2014.

[12] Krizhevsky A, Sutskever I, Hinton G E.ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems,2012:1097-1105.

[13] 薛燕娜.机器学习算法在蛋白质结构预测中的应用[D].江苏无锡:江南大学,2016.

[14] Cortes C, Vapnik V.Support-vector networks[M].[S.l.]: Kluwer Academic Publishers,1995.

[15] Chang C C, Lin C J.LIBSVM: A library for support vector machines[M].[S.l.]: ACM,2011.

[16] Verma R, Razis S.Medical- relation- extraction[EB/OL].[2017-08-07].https://github.com/ammskang/Medical-Relation-Extraction.

[17] Buckland M, Gey F.The relationship between recall and precision[J].Journal of the Association for Information Science & Technology,1994,45(1):12-19.

[18] Flores F N, Moreira V P, Heuser C A.Assessing the impact of stemming accuracy on information retrieval[C]// International Conference on Computational Processing of the Portuguese Language,2010:11-20.

[19] Lin Y, Shen S, Liu Z, et al.Neural relation extraction with selective attention over instances[C]//Meeting of the Association for Computational Linguistics,2016:2124-2133.

[20] Cohen T, Widdows D.Empirical distributional semantics: Methods and biomedical applications[J].Journal of Biomedical Informatics,2009,42(2):390-405.

计算机工程与应用 www.cai.org