doi:10.6041/j.issn.1000-1298.2018.12.026

基于卷积模型的农业问答语性特征抽取分析

张明岳[1]　吴华瑞[1,2]　朱华吉[2,3]

(1.国家农业信息化工程技术研究中心,北京 100097;2.北京农业信息技术研究中心,北京 100097;
3.农业农村部农业信息软硬件产品质量检测重点实验室,北京 100097)

摘要:互联网农技推广社区每秒增衍问答数据近万组,这些海量数据具有隐性的词性、情感和冗余向量特征,实现数据聚合与数据块消减是该领域的难题。提出了一种基于卷积神经网络的农业问答情感极性特征抽取分析模型,结合农业分词字典,对数据集进行分词后使用 Skip-gram 模型转换为 256 维的词向量,利用批规范后的卷积神经网络对数据集进行训练,从而得到用于识别农技推广社区问答词性情感相似性的神经网络模型参数。试验结果表明,该方法能够准确识别测试样例集中的冗余队列,与其他 5 种文本分类方法进行比较,各项指标优势明显,针对测试集的语性特征抽取准确率达到 82.7%。

关键词:农业信息分类;特征提取;自然语言处理;卷积神经网络

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1000-1298(2018)12-0203-08

Analysis of Extraction of Semantic Feature in Agricultural Question and Answer Based on Convolutional Model

ZHANG Mingyue[1]　WU Huarui[1,2]　ZHU Huaji[2,3]

(1. *National Engineering Research Center for Information Technology in Agriculture*, *Beijing* 100097, *China*
2. *Beijing Research Center for Information Technology in Agriculture*, *Beijing* 100097, *China*
3. *Key Laboratory of Agricultural Information Software and Hardware Product Quality Testing*, *Ministry of Agriculture and Rural Affairs*, *Beijing* 100097, *China*)

Abstract: Tens of thousands of question and answer data have been increased per second in the internet agricultural technology extension community, these massive data have features of recessive part of speech, emotion and unwanted vectors, and how to implement data aggregation and data block reduction is the difficult problem in this field. An analytical model for the extraction of emotional polarity in agricultural question and answer based on convolutional neural network was proposed, the training set was transformed into a 256-dimensional word vector by using the Skip-gram model after segmenting the dataset with agricultural word segmentation dictionary. The convolution neural network after batch-normalization specification was used to train the dataset, and the neural network model parameters used to identify the part of speech emotional similarities in the agricultural technology promotion community question and answer were obtained. The experimental results showed that the method could accurately identify redundant queues in the test sample set, and by comparing with the other four text classification methods, there were also obvious advantages in each index, the accuracy of the semantic feature extraction for the test set was up to 82.7%.

Key words: classification of agriculture information; feature extraction; natural language processing; convolutional neural network

0　引言

问答社区是基于互联网,以用户提出问题、回答问题和讨论问题为主的知识服务社区,能够更好地满足互联网用户获取信息和交流知识的需求,是目前自然语言处理(Natural language processing, NLP)

收稿日期:2018-05-23　修回日期:2018-09-03

基金项目:国家自然科学基金项目(61571051)、北京市自然科学基金项目(4172024)和北京市农林科学院 2018 年度科研创新平台建设项目(PT2018-25)

作者简介:张明岳(1989—),男,研究实习员,主要从事农业智能系统及数据挖掘研究,E-mail:zhangmy@nercita.org.cn

通信作者:吴华瑞(1975—),男,研究员,主要从事农业智能系统研究,E-mail:wuhr@nercita.org.cn

和信息检索(Information retrieval,IR)领域备受关注、具有广泛发展前景的研究方向[1-2]。“中国农技推广 APP”作为服务于农技人员的专业平台,用户每天在农技问答模块发布的提问有上万余条,这类文本具有稀疏性、实时性、不规范等特点,加剧了问题文本关键词特征的稀疏化,难以充分挖掘特征之间的关联性,如何从数据集中方便、快捷地挖掘有效信息并提供更高质量和智能化的农业信息服务已成为农业信息分类领域文本挖掘的主要任务之一。传统的人工筛查需要消耗大量的人力、物力,并且很难高效地完成对无效冗余数据的处理。目前常用的人工特征分类及浅层分类学习模型虽然能够辅助完成数据筛查及剔除等工作,但由于其过分依赖人工选取特征和分类器性能,不具备从数据中自动抽取和组织信息的能力,导致经典的文本分析方法在短文本处理上的适用性下降[3-4]。因此利用计算机实现农技冗余问答自动、智能筛查是“中国农技推广 APP”需要解决的一个重要问题。神经网络模型具有灵活性和多样性的特点,在序列标注[5]、语义匹配[6]、情感分析[7]等自然语言处理任务中表现出较好的性能,由于该类模型能够以端到端的方式进行训练,自动学习特定任务并挖掘文本内的大量语义关系,有效减少了传统的统计机器学习方法中人工设定大量特征等相关工作[8]。

目前结合神经网络模型开展自然语言处理的相关应用已经取得了一定成果,其中卷积神经网络(Convolutional neural network,CNN)在情感分析和文本分类领域得到很好的应用[9-12]。由于农业领域一直缺乏大规模可用的数据库,因此关于这方面的研究还较少,只有个别研究者针对农业特定领域研究神经网络模型在农业问答系统的应用,但仍处于起步阶段。赵明等[13]构建了基于 Word2vec 和双向门控循环单元神经网络(Bi-directional gated recurrent unit,BIGRU)的番茄病虫害问句分类模型,对番茄病虫害智能问答系统用户问句进行高效分类。针对传统的句子相似度算法准确率较低的问题,梁敬东等[14]通过构建基于 Word2vec 和长短期记忆网络(Long short-term memory,LSTM)的神经网络计算问句相似度,并在水稻常问问题集(Frequently asked question,FAQ)中的问句上进行验证,以提高系统回答的准确性。以上研究的开展为神经网络应用于农业知识问答系统提供了参考和可行性依据,但关于神经网络应用于文本多样性、情感极性等农业文本特征挖掘方面仍有不足,关于利用卷积神经网络检验农技推广提问数据的精确性和可靠性方面尚未见报道。

为了实现农技推广社区问答情感特征信息的有效挖掘和表达,本文利用基于卷积神经网络模型的知识自动化的方法,有针对性地引入农业词库字典进行中文分词和词向量表示[15],利用卷积神经网络提取文本情感表达作为文本特征向量,用于情感分类,并进一步针对其重要的结构参数和训练策略进行优化和改进,构建一种基于卷积神经网络的农业问答情感极性特征抽取分析模型,以实现农技推广提问的精确高效识别。

1 数据采集与预处理

1.1 样本采集

本文数据集来源于“中国农技推广 APP”农技问答模块,以 2017 年 8 月上线到 2018 年 4 月产生的 130 多万条提问数据作为基础样本。由于人工标注百万级样本十分困难,参照文献[16-18]使用的文本分类数据集量级,根据月份选取 8 000 条数据作为试验样本集。其中人工标注有效及无效提问各 3 000 条作为学习数据集,用于卷积神经网络训练和优化参数验证,人工选择样例如表 1 所示。剩余 2 000 条样本数据作为模型效果验证的测试集,由于测试集是在训练集和验证集选取之后选取,已经较大限度地保证了训练与测试数据集文本的不重叠,因此可以将测试结果的平均准确率作为文本模型的识别效果评价指标[19]。

表 1 人工选择样例

Tab.1 Worked examples of manual annotation

编号	农技提问文本	人工标注
1	豇豆的高产栽培管理技术?	有效
2	这个柑橘树怎么了? 如何防治?	有效
3	如何选择优质油菜品种?	有效
4	大米加工后,如何选择优质米?	有效
5	这是什么虫子? 用什么药防治?	有效
6	如何防治水稻二化螟?	有效
7	求慈姑的种植方法及管理技术?	有效
8	在鲜花似锦的季节里,知道它的花名吗?	无效
9	你猜一猜这是什么水果的树?	无效
10	请大家猜猜这是什么植物?	无效
11	请问这是什么,你吃过吗?	无效
12	我家的盆景石榴漂亮吗?	无效
13	这种小区环境居住怎样?	无效
14	丝瓜怎样烹饪好吃? 是否要去皮?	无效

1.2 数据集预处理

中文文本需要进行预处理转换为数字形式,以便能够被计算机识别。为最大程度地保留原始中文文本的特征及语义信息,减少信息损失,需要对文本进行去噪、分词、向量表示等预处理操作,主要步骤

如图 1 所示。

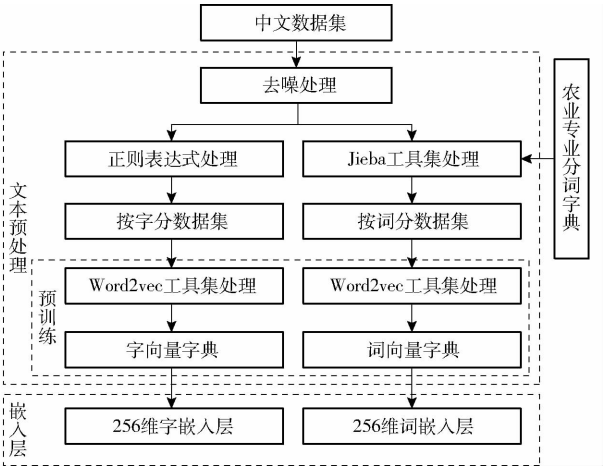

图 1　数据预处理过程示意图

Fig. 1　Schematic of data preprocessing

(1)去噪:数据集中原始数据包含中文特殊字符、英文特殊字符、空格等多种类型的符号信息,不利于语性特征抽取。因此使用正则表达式对数据集进行去噪处理,仅保留中文、英文、字母、数字等通用特征信息。

(2)分字与分词:利用 Python 正则表达式对数据集中每条语句的汉字进行分割形成分字数据集。由于中文分词[20]主要依赖语义与语境,而农技提问又包含很多农业专业词汇,基础分词库很难满足要求,本试验还需要建立农业专业词汇的自定义分词字典。参照文献[21]选择搜狗农业词汇大全中的 8 874 个词汇作为农业专业分词字典,再利用 Jieba 分词工具包对数据集进行精确模式分词形成分词数据集[22]。

(3)生成词向量:使用 Word2vec 工具集的 Skip-gram 模型[23]对分字集和分词集进行预训练,具体操作方法是对文本中的字、词等元素的出现频率进行统计,通过无监督训练,获得作为语料基础构成元素的字、词对应的指定维度的向量表征,最终生成指定维度的字向量和词向量。

(4)文本向量化:为便于神经网络训练,文本数据需要转化为字或词嵌入,具体操作是将样本中每个字或词替换成对应的向量表示,将文本转化为向量组。对样本每条数据的字或词进行统计,选择字或词数最多的那条文本的字或词个数作为文本向量维度,其余提问长度不足的通过 0 来补齐。

2　基于卷积神经网络模型的农业数据筛查方法

自 KIM[10]研究了利用卷积神经网络处理自然语言后,大量研究人员在其基础上做了拓展与优化,尽管文本分类模型变得越来越丰富,但所有模型的基本架构都与图 2 相近。基本思路是字或词经过嵌入层后利用不同神经网络结构提取局部、全局和上下文信息,经过全连接层合并到一起,最后利用不同分类器进行文本分类得到结果。

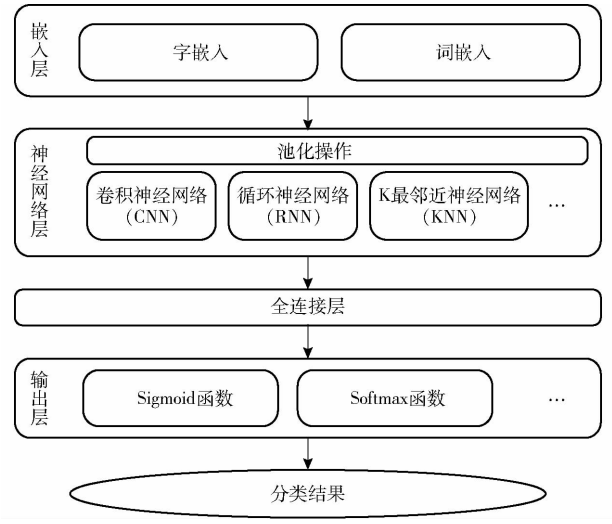

图 2　文本分类基本架构

Fig. 2　Basic structure of text categorization

本文在基本架构基础上进行了拓展,增加了卷积层数以及更多尺度的卷积核,同时在激活之前增加了批标准化进行规范化处理,全连接层中也增加了批标准化处理,最后使用 Softmax 逻辑回归作为分类器,进行数据的语性特征抽取。

模型中卷积核的尺寸与数量对于 CNN 的性能至关重要。输入语料通过 i 个不同的卷积核卷积,生成 j 个不同的特征图,卷积层满足公式

$$x_j^{(k)} = f\left(\sum_{i \in M_j} x_i^{(k-1)} W_{ij}^{(k)} + b_j^{(k)} \right) \tag{1}$$

式中　$x_j^{(k)}$——在 k 卷积层的第 j 个特征图

$f(\cdot)$——批标准化及激活函数

M_j——输入图像的特征量

$W_{ij}^{(k)}$——卷积核

$b_j^{(k)}$——偏置值

针对各层分布不均和精度弥散等问题,在进行激活之前使用批标准化(Batch normalization,BN)来规范响应,同时加快网络收敛,防止过拟合。具体公式为

$$\mu_{\mathrm{B}} = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{2}$$

$$\sigma_{\mathrm{B}}^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_{\mathrm{B}})^2 \tag{3}$$

$$\hat{\chi}_i = \frac{x_i - \mu_{\mathrm{B}}}{\sqrt{\sigma_{\mathrm{B}}^2 + \varepsilon}} \tag{4}$$

$$y_i = \gamma \hat{\chi}_i + \beta = \mathrm{BN}_{\gamma,\beta}(x_i) \tag{5}$$

式中 x——输入值

m——批量化的数目

γ、β——学习参数

μ_B——均值 σ_B^2——方差

$\hat{\chi}_i$——输入值归一化值

ε——常量,用来保证值的稳定性

y_i——结果输出值

$BN_{\gamma,\beta}(\cdot)$——批标准化函数

参考 CLEVERT 等[24]的试验,模型激活函数使用修正线性单元(Rectified linear unit, ReLU),公式为

$$f(x) = \max(0, x) \tag{6}$$

根据文本分类的特性,需要在一定程度上降低卷积层参数误差造成的估计均值偏移所引起的特征提取的误差,试验选用 Max-pooling 作为池化方法。网络的训练阶段使用批量随机梯度下降法(Mini-batch stochastic gradient descend)。

本文使用 Softmax 逻辑回归来做特征分类器(对应 Softmax loss 损失函数),进行实际文本的语性特征抽取[25]。最终确定的卷积神经网络结构如图 3 所示。

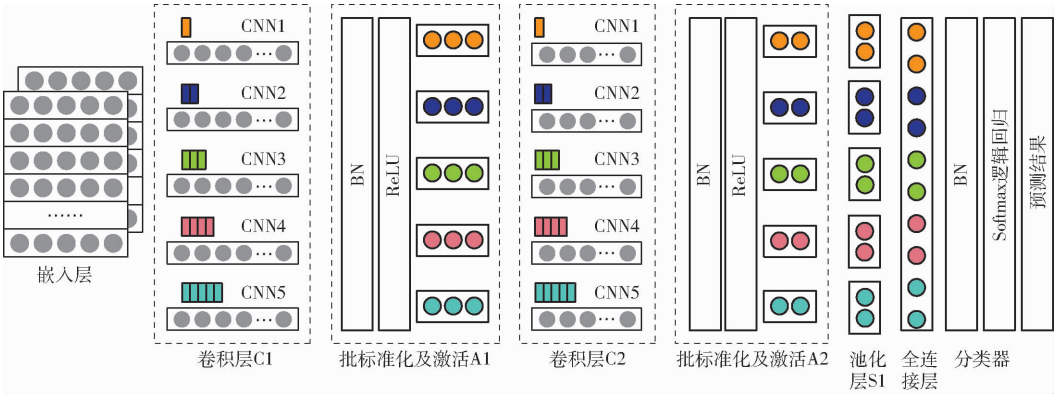

图 3 基于文本的卷积神经网络结构示意图

Fig. 3 Schematic of text-based convolution neural network

3 试验及结果分析

3.1 硬件及软件

本试验处理平台为联想台式计算机,处理器为 Intel(R) Core(TM) i5 - 4590、主频 3.30 GHz、内存 8 GB、容量 120 GB 金士顿固态硬盘,运行环境为:Windows 10 专业版 64 位,软件环境为 Python 3.6.5 和 Tensorflow 1.8.0。

3.2 试验操作流程

(1)输入层

输入层为经过预处理的 256 维词嵌入,对分词后数据集的词组个数进行统计,可以得到数据集中最多词数为 58 个,即每条提问的词向量维度为 58 × 256。将输入数据顺序打乱并随机排列,选取前面 90%(5 400 条)作为训练数据,后面 10%(600 条)作为验证数据。训练次数设置为 200 次,每批次输入 500 条,共计输入 2 200 批次,图 4 为输入层中“植物”一词的向量示例。

(2)卷积层

卷积层的作用是特征提取,设置卷积核长度为 58,窗口层数为 5,每层窗口滑动尺寸分别是 1 ~ 5,卷积核每个窗口特征映射数为 200,所以第 1 层卷积核 W1 的尺寸为(1 ~ 5) × 58 × 1 × 200,第 2 层卷积核 W2 的尺寸为(1 ~ 5) × 58 × 200 × 200。

(3)池化层

池化层的作用是特征压缩,在进行池化前使用了批标准化进行处理。最后连接一个 Softmax 逻辑回归分类器,用于将压缩好的特征映射到输出层。S1 对前面的特征图进行了最大池化操作,每批次得到 500 个 1 000 维的特征图。

(4)Softmax 分类器

经过训练,最后剩下的神经元由 Softmax 分类器将其拼合成为一维列向量,全连接到输出层,计算出属于每类特征输出的概率值。

(5)输出层

比较分类器中计算出的语性特征概率值,将结果归类到概率最大的一组,然后合并归类结果并保存到 prediction.csv 文件中,识别结果样例如表 2 所示,表中“○”表示识别结果与真实值相同,“X”表示识别结果与真实值不同。

3.3 结果与误差分析

采用试验所描述的卷积神经网络结构使用训练样本来训练模型,网络权重初始化采用标准差为 0.01、均值为 0 的高斯分布,样本迭代次数均设置为 200,批处理尺寸设置为 100,设置权重参数的初始学习速率为 0.001,动量因子设置为 0.9。对上述训练集做 2 600 次迭代训练,其训练曲线如图 5 所示。

从图 5 可以看出,随着迭代次数不断增加,模型

植物
[3.10529664e-04 4.09875269e-04 -1.58788776e-03 2.28298432e-03 3.27867776e-04 9.27895133e-04
 -3.07425763e-03 2.16609705e-03 5.68162824e-04 1.70148944e-03 -5.99219347e-04 -1.83012255e-03
 -5.69412892e-04 2.39426713e-03 2.01639417e-03 8.36299674e-04 1.39208534e-03 2.09972658e-03
  3.56778945e-03 9.36444412e-05 2.47261883e-03 1.86652120e-04 8.35239131e-04 1.91353925e-03
  1.33107894e-03 -4.63300501e-04 2.12889188e-03 3.80182988e-03 2.32619862e-03 -1.93791115e-03
  1.96267315e-03 9.89724576e-05 5.43027942e-04 1.91218092e-03 -9.87640815e-04 2.70729419e-03
 -5.22366201e-04 -2.30939360e-03 -7.23017787e-04 2.62568443e-04 1.63415831e-03 1.13233994e-03
  2.93529010e-03 1.55370857e-03 -5.01632858e-05 -1.02271978e-03 3.41642415e-04 1.07262644e-03
  1.10205286e-03 1.26316596e-03 -1.34182372e-03 3.32624838e-03 -1.85069127e-03 -6.39962207e-04
  1.36651983e-03 1.10262434e-03 1.45157811e-03 7.82271905e-04 1.35596364e-03 1.09340355e-03
  1.40537985e-03 9.52932751e-04 -5.41562506e-04 -1.37049856e-03 -6.11570373e-04 3.14153335e-03
  9.07721696e-04 -2.13261601e-03 3.86598294e-05 -6.61627797e-04 2.25226395e-03 -3.42257461e-04
 -1.26645697e-04 2.35449383e-03 6.07525173e-04 3.24225938e-03 -2.88425782e-03 4.68967424e-04
 -1.10580004e-03 4.08961717e-03 -3.96234216e-04 4.84531804e-04 -1.82612217e-03 -1.25477812e-03
 -2.38927733e-03 2.73740388e-05 9.34268057e-04 -2.56357162e-04 1.22177426e-03 -2.67454865e-03
 -1.06755528e-03 -3.53372027e-03 2.31725600e-04 3.50663468e-04 -9.03385633e-04 1.69165828e-03
 -2.24331068e-03 -5.94433986e-05 -9.96161485e-04 -7.95818574e-04 -3.81366175e-04 8.22727336e-04
  1.93403021e-03 1.31992355e-03 4.59108160e-05 -2.57842862e-06 -2.23831856e-03 1.73781021e-03
  2.48707540e-04 1.46664924e-03 3.38902784e-04 -8.74277845e-04 1.99958053e-03 -1.19238452e-03
 -1.00566738e-03 -1.56576186e-03 -8.07943288e-04 2.50367634e-03 1.12716574e-04 2.38987198e-03
 -2.44226400e-03 -1.09547062e-03 2.39005173e-03 8.01117800e-04 -1.83232618e-03 1.07374438e-03
 -2.77022715e-04 4.80764807e-04 -1.78755872e-04 1.02038775e-03 -3.49068723e-04 -5.06368640e-04
  1.05727243e-03 6.55365351e-04 -1.02271850e-03 3.48501810e-04 -1.25083397e-03 2.50973622e-03
  1.04220305e-03 9.64761653e-04 -1.51284889e-03 -1.43137353e-03 -2.64626462e-03 -1.64428842e-04
  3.85076040e-04 1.72254746e-03 -2.55771424e-03 2.75811204e-03 -1.32115977e-03 2.17787549e-03
  2.05950509e-03 -1.51204003e-03 1.99999427e-04 3.06186872e-03 -4.92122490e-03 3.13210505e-04
  9.84701561e-04 1.79583614e-04 5.72731078e-04 -2.25051306e-03 -6.18989929e-04 -2.45922734e-03
  2.49257032e-03 2.75666011e-03 -3.78999393e-03 -2.60237284e-05 1.48615148e-03 -1.27093296e-03
  2.55148771e-04 6.72247319e-04 1.73833148e-04 1.89925544e-03 3.94173944e-03 4.02001315e-04
  7.56311783e-05 -1.15518423e-03 1.91788145e-04 -1.18543947e-04 -1.00238982e-03 1.96692231e-03
 -5.18608780e-04 -1.16999820e-03 1.92416855e-03 9.49152105e-04 -8.29774130e-04 -8.27991869e-04
 -1.29414874e-03 2.10061320e-03 7.37149385e-04 8.27086158e-04 1.60895591e-03 2.79979175e-03
 -1.96671463e-03 8.78009741e-05 -2.33619520e-03 -4.76018351e-04 -1.24412865e-04 -1.25177903e-03
 -4.52481298e-04 -5.05762873e-04 2.50766100e-03 -2.13162508e-03 -9.84595157e-04 -1.35231880e-03
 -9.64650535e-04 1.65500969e-03 -3.37862846e-04 1.28140126e-03 6.28238660e-04 5.40075125e-04
  3.72687122e-04 1.65008462e-03 1.70217862e-03 1.80542865e-03 3.81135091e-04 1.64178968e-03
  2.31536556e-04 1.77176890e-03 3.53876472e-04 -9.02965316e-04 1.40889979e-03 4.66602592e-04
  9.16808727e-04 -4.14801354e-04 2.88227893e-04 -8.23142182e-04 1.21107046e-03 -7.94069842e-04
 -3.18893726e-04 1.04034541e-03 6.13024109e-04 -3.85895284e-04 1.72666309e-03 -6.77483506e-04
 -1.15147699e-03 -1.85773312e-03 8.08102312e-04 2.15630699e-03 3.61727946e-03 1.46073441e-03
  8.91670876e-04 1.52314373e-03 -2.39899289e-03 5.71614655e-04 -1.93693815e-03 -8.29574361e-04
  1.61845132e-03 6.92240021e-04 9.45764012e-04 4.39916388e-04 -1.26754632e-03 1.39096705e-03
  1.63663737e-03 8.48787196e-04 -2.21363292e-03 1.16078544e-03]

图 4　词向量示例

Fig. 4　Examples of word vectors

表 2　模型识别结果样例

Tab. 2　Worked examples of model identification results

农技提问文本	识别结果	真实值	校验结果
“请问 玉米 得 了 什么 病”	有效	有效	○
“杜鹃花 常见 病虫害 如何 防治”	有效	有效	○
“杜鹃花 栽培 管护 要点 有 哪些”	有效	有效	○
“这是 什么 作物 的 果实”	有效	有效	○
“土蜂蜜 有 什么 药用价值”	有效	有效	○
“韭 叶 兰花 漂亮 吧”	有效	无效	X
“请问 这是 什么 花 漂 不 漂亮”	有效	无效	X
“这是 什么 作物 什么 季节 煮汤 喝好”	无效	无效	○
“玫瑰花 咋样 啊”	无效	无效	○
“这是 什么 水果 好吃 吗”	无效	无效	○
“荷花 怎么 高出 水面 那么 多”	无效	无效	○
“这 又 是 什么 花 绿意盎然 充满生机”	无效	无效	○
“猪用 防疫 分 哪几种 怎样 贮存”	无效	有效	X
“为什么 苹果 的 品种 要 合理 搭配”	无效	有效	X

分类误差逐渐降低。当训练迭代到 2 000 次时训练集的识别准确率最高达到 98.6%，迭代到 2 200 次时验证集的识别率最高达到 93.5%，且从第 1 400 次迭代以后训练集和验证集两者的误差差值趋于稳

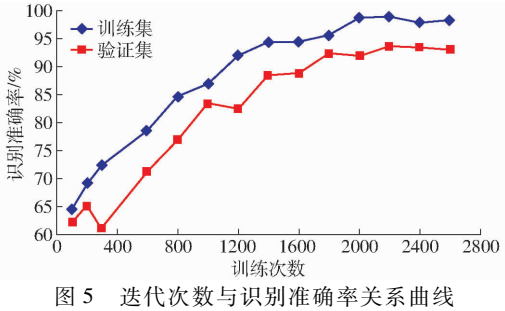

图 5　迭代次数与识别准确率关系曲线

Fig. 5　Diagram of relationship between number of iterations and accuracy

定，说明模型状况良好，卷积神经网络达到了预期的训练效果。由试验可以确定训练达到 2 200 次以后模型对样本的识别准确率趋于拟合，将训练次数设定为 2 200 能够使模型得到充分训练。

为了验证不同类型嵌入层对模型效果的影响，分别使用字向量嵌入、词向量嵌入以及经过农业字典分词的词向量嵌入作为输入层，对试验模型进行 2 200 次的迭代对比训练，识别结果如图 6 所示。

由图 6 可以看出，随着迭代次数增加，各模型识别准确率均不同程度增加，当上涨到一定程度后各模型识别率趋于稳定。经过 2 200 次训练，词向量

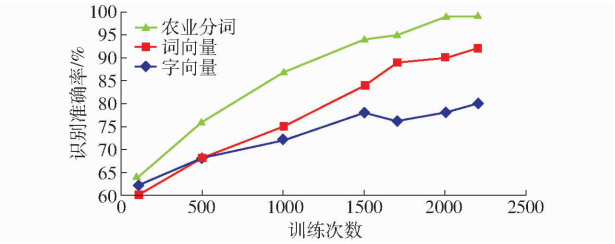

图 6 不同嵌入层迭代次数与识别准确率关系曲线

Fig. 6 Diagram of relationship between number of iterations and accuracy in different embedded layers

嵌入的识别准确率最高达到 92%，字向量嵌入的识别准确率最高达到 80%，经过农业字典分词的向量嵌入识别准确率是三者中最高的，接近 99%。试验证明，输入层使用分词嵌入能够比分字更好地表达文本特征，针对所属领域使用专用的词汇进行细化分词后会更加充分地表达文本特征。

通过表 3 可以看出，增加卷积核滑动窗口个数以及窗口特征映射层数能够有效增加模型的识别准确率。当模型参数增加到一定程度后继续增加参数宽度和深度，模型的识别准确率很难继续提升，但模型需要的训练时间更长。通过模型参数比较，设定卷积核的窗口宽度为 5、映射特征层数为 200 的训练模型能够在现有软硬件的条件下较好地满足试验要求。

表 3 试验模型参数的比较

Tab. 3 Comparison of experimental model parameters

窗口数	窗口宽度	窗口特征层数	训练时间/s	识别准确率/%
3	3,4,5	200	1 945.7	81.4
4	2,3,4,5	200	2 343.8	82.5
5	1,2,3,4,5	50	945.6	70.0
5	1,2,3,4,5	100	1 493.0	79.3
5	1,2,3,4,5	200	2 926.8	82.7
6	1,2,3,4,5,6	250	5 121.3	82.8

为了进一步证明提出方法的性能，将其与现有的 JOHNSON 等[26]提出的 One-hot 词表示法 + CNN 的文本分类方法、ASEERVATHAM 等[27]提出的 SVM 分类器方法、DANTI 等[28]提出的文档矢量空间表示模型(DVSM) + 词间距离度量分类方法、ZHANG 等[29]提出的 KNN 分类器方法以及使用 Dropout 代替 Batch - Normalization 执行标准化的 CNN 分类方法等 5 种文本分类方法进行比较，对测试集的 2 000 条提问数据进行识别，各种分类方法的筛选性能如表 4 所示。

通过表 4 可以看出，各类算法都能够对测试集进行有效的特征筛选，本文使用方法在 6 种算法中识别准确率最高，达到了 82.7%。尽管文献[26]的方法也使用了 CNN 模型，但由于输入层使用的是

表 4 各种分类方法的比较

Tab. 4 Comparison of various classification methods

算法	精确率/%	召回率/%	F1 度量值	识别准确率/%	训练时间/s
One-hot + CNN[26]	81.6	78.9	80.2	68.2	3 238.8
SVM[27]	80.3	77.2	78.7	73.1	807.0
DVSM + 距离度量[28]	84.9	81.4	83.1	72.0	931.2
KNN[29]	89.2	82.9	85.9	78.8	991.6
Dorpout + CNN	89.0	82.9	85.8	76.4	1 950.4
本文方法	92.1	87.5	89.7	82.7	2 926.8

One-hot 方法，其准确率只达到 68.2%，明显低于其他筛选方法，说明 Word2vec 的 Skip - gram 模型能够更高效地表示语料特征，也证明了输入层的文本处理方式对于模型训练结果存在较大影响。虽然文献[27]方法在测试集中的识别准确率比文献[28]方法高出 1.1 个百分点，但是精确率和 F1 度量值明显不如后者，这也间接说明相邻分词之间的关联语意对识别结果存在影响。使用 Batch - Normalization 规范响应相较于卷积神经网络常用的 Dorpout 标准化方法能够加快收敛，使训练更加充分，防止过拟合，显著提高识别准确率，识别准确率高出 6.3 个百分点。综合表 4 列出的各类文本分类方法，本文提出的基于 CNN 优化模型因为权值共享机制减少了网络中的可训练参数，有效降低了模型复杂度，具有更好的泛化能力，因此相较于其他机器学习模型取得了更好的分类效果[30]。卷积神经网络的核心特点是每个卷积层包含数个卷积核及大量特征面，通过池化操作大量减少模型中的神经元个数，增强了模型表达能力，因此对输入空间的平移不变特征更具鲁棒性，有效防止训练过拟合[31]。尽管卷积神经网络模型训练时间远高于表 4 其他分类方法，但通过权值共享、局部连接、批标准化增强、池化操作等使本文方法具有更少的连接和参数、更易于训练，具有自动抽取语性特征并且得到更多分类特征的特点。

4 结论

(1)研究方法满足实际应用需求。通过卷积网络模型筛选数据，减小了人工筛查的工作强度，避免了传统识别方法中复杂的预处理和特征筛选过程，提高了算法优化效率，对测试集特征识别准确率达到 82.7%。

(2)优化输入层表示及模型结构能显著提高识别效果。不同类型嵌入层对于筛选结果也有较大影响，使用农业专业词典进行分词处理的嵌入层在模型学习效率和识别准确率上都有提高。另外使用

Batch－Normalization 替换 Dropout 训练后识别效果相较于 Dropout 标准化的卷积神经网络识别准确率提升了 6.3 个百分点，对比其他类型的文本分类模型识别效果也具有明显优势。

参 考 文 献

1 袁毅，杨莉. 问答社区用户生成资源行为及影响因素分析：以百度知道为例[J]. 图书情报工作，2017，61(22)：20－26.
YUAN Yi，YANG Li. Users' behavior of generating resources and influence factors in the Q&A community：evidence from Baidu Knowns[J]. Library and Information Service，2017，61(22)：20－26.（in Chinese）

2 WANG B，WANG X，SUN C，et al. Modeling semantic relevance for question-answer pairs in web social communities[C]// Meeting of the Association for Computational Linguistics. Cambridge：Association for Computational Linguistics，2010：1230－1238.

3 罗世奇，田生伟，孙华，等. 栈式自编码的恶意代码分类算法研究[J]. 计算机应用研究，2018，35(1)：56.
LUO Shiqi，TIAN Shengwei，SUN Hua，et al. Research on malicious code classification algorithm of stacked auto encoder[J]. Application Research of Computers，2018，35(1)：56.（in Chinese）

4 奚雪峰，周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报，2016，42(10)：1445－1465.
XI Xuefeng，ZHOU Guodong. A survey on deep learning for natural language processing[J]. Acta Automatica Sinica，2016，42(10)：1445－1465.（in Chinese）

5 COLLOBERT R，WESTON J，BOTTOU L，et al. Natural language processing（almost）from scratch[J]. The Journal of Machine Learning Research，2011，12：2493－2537.

6 金丽娇，傅云斌，董启文，等. 基于卷积神经网络的自动问答[J]. 华东师范大学学报(自然科学版)，2017(5)：66－79.
JIN Lijiao，FU Yunbin，DONG Qiwen，et al. The auto-question answering system based on convolution neural network[J]. Journal of East China Normal University（Natural Sciences），2017(5)：66－79.（in Chinese）

7 PORIA S，CAMBRIA E，GELBUKH A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing，2015：2539－2544.

8 QIU X，HUANG X. Convolutional neural tensor network architecture for community-based question answering[C]// International Conference on Artificial Intelligence，2015：1305－1311.

9 KALCHBRENNER N，GREFENSTETTE E，BLUNSOM P. A convolutional neural network for modelling sentences[J/OL]. arXiv preprint arXiv：1404.2188，2014，1.

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

11 WEHRMANN J，BECKER W，CAGNINI H E L，et al. A character-based convolutional neural network for language-agnostic Twitter sentiment analysis[C]// Neural Networks（IJCNN），2017 International Joint Conference on. IEEE，2017：2384－2391.

12 邓憧. 基于 CNN 语义匹配的自动问答系统构建方法研究[D]. 哈尔滨：哈尔滨工业大学，2017.
DENG Chong. Research on constructing automatic question answers system based on convolution neural network of semantic matching[D]. Harbin：Harbin Institute of Technology，2017.（in Chinese）

13 赵明，董翠翠，董乔雪，等. 基于 BIGRU 的番茄病虫害问答系统问句分类研究[J/OL]. 农业机械学报，2018，49(5)：271－276. http：//www.j-csam.org/jcsam/ch/reader/view_abstract.aspx？flag=1&file_no=20180532&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2018.05.032.
ZHAO Ming，DONG Cuicui，DONG Qiaoxue，et al. Question classification of tomato pests and diseases question answering system based on BIGRU[J/OL]. Transactions of the Chinese Society for Agricultural Machinery，2018，49(5)：271－276.（in Chinese）

14 梁敬东，崔丙剑，姜海燕，等. 基于 word2vec 和 LSTM 的句子相似度计算及其在水稻 FAQ 问答系统中的应用[J]. 南京农业大学学报，2018，41(5)：946－953.
LIANG Jingdong，CUI Bingjian，JIANG Haiyan，et al. Sentence similarity computing based on word2vec and LSTM and its application in rice FAQ question-answering system[J]. Journal of Nanjing Agricultural University，2018，41(5)：946－953.（in Chinese）

15 LAI S，LIU K，XU L，et al. How to generate a good word embedding？[J]. IEEE Intelligent Systems，2016，31(6)：5－14.

16 PANG B，LEE L. A sentimental education：sentiment analysis using subjectivity summarization based on minimum cuts[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics，2004：271.

17 PANG B，LEE L. Seeing stars：exploiting class relationships for sentiment categorization with respect to rating scales[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics，2005：115－124.

18 LI X，ROTH D. Learning question classifiers[C]// Proceedings of the 19th International Conference on Computational Linguistics—Volume 1. Association for Computational Linguistics，2002：1－7.

19 杨国国，鲍一丹，刘子毅. 基于图像显著性分析与卷积神经网络的茶园害虫定位与识别[J]. 农业工程学报，2017，33(6)：156－162.
YANG Guoguo，BAO Yidan，LIU Ziyi. Localization and recognition of pests in tea plantation based on image saliency analysis and convolutional neural network[J]. Transactions of the CSAE，2017，33(6)：156－162.（in Chinese）

20 CHEN X, QIU X, ZHU C, et al. Long short-term memory neural networks for chinese word segmentation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1197 - 1206.
21 赵静. 大规模汉语语义词典构建[D]. 哈尔滨:哈尔滨工业大学, 2011.
ZHAO Jing. Building a large scale Chinese semantic dictionary[D]. Harbin: Harbin Institute of Technology, 2011. (in Chinese)
22 YAN Y, WANG C, SHI W. Survey of researches on Chinese sentiment analysis based on deep learning[C]// 3rd International Conference on Artificial Intelligence and Industrial Engineering, Shanghai, 2017.
23 KIM A Y, HA J G, CHOI H, et al. Automated text analysis based on skip-gram model for food evaluation in predicting consumer acceptance[J]. Computational Intelligence and Neuroscience, 2018(2): 9293437.
24 CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (elus)[J/OL]. arXiv preprint arXiv: 1511.07289, 2015.
25 CUI Y, ZH F, LIN Y, et al. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop[C]// Computer Vision and Pattern Recognition. IEEE, 2016: 1153 - 1162.
26 JOHNSON R, ZHANG T. Effective use of word order for text categorization with convolutional neural networks[J/OL]. arXiv preprint arXiv: 1412.1058, 2014.
27 ASEERVATHAM S, ANTONIADIS A, GAUSSIER E, et al. A sparse version of the ridge logistic regression for large-scale text categorization[J]. Pattern Recognition Letters, 2011, 32(2): 101 - 106.
28 DANTI A, BHUSHAN S. Document vector space representation model for automatic text classification[C]// Proceedings of International Conference on Multimedia Processing, Communication and Information Technology, 2013: 338 - 344.
29 ZHANG S, LI X, ZONG M, et al. Learning K for KNN classification[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2017, 8(3): 43.
30 HUANG J T, LI J, GONG Y. An analysis of convolutional neural networks for speech recognition[C]// Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015: 4989 - 4993.
31 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229 - 1251.
ZHOU Feiyan, JIN Linpeng, DONG Jun. Review of convolutional neural network[J]. Chinese Journal of Computers, 2017, 40(6): 1229 - 1251. (in Chinese)

(上接第 194 页)

13 BAZI Y, MALEK S, ALAJLAN N, et al. An automatic approach for palm tree counting in UAV images [C]// Geoscience and Remote Sensing Symposium. IEEE, 2014: 537 - 540.
14 MORANDUZZO T, MELGANI F. Automatic car counting method for unmanned aerial vehicle images [J]. IEEE Transactions on Geoscience & Remote Sensing, 2014, 52(3): 1635 - 1647.
15 STEVEN W C, SHREYAS S S, SANDEEP D, et al. Counting apples and oranges with deep learning: a data driven approach [J]. IEEE Robotics & Automation Letters, 2017, 2(2): 781 - 788.
16 KAMILARIS A, PRENAFETA-BOLDÚ F X. Deep learning in agriculture: a survey [J]. Computers & Electronics in Agriculture, 2018, 147(1): 70 - 90.
17 ADRIAN C, CARLOS S, ALEJANDRO R R, et al. A review of deep learning methods and applications for unmanned aerial vehicles [J]. Journal of Sensors, 2017(2): 1 - 13.
18 ANDRES M, PHILIPP L, CYRILL S. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns [J/OL]. [2017 - 09 - 20]. https://arxiv.org/abs/1709.06764.
19 PHILIPP L, MARKUS H, SLAWOMIR S, et al. Effective vision-based classification for separating sugar beets and weeds for precision farming [J]. Journal of Field Robotics, 2017, 34(6): 1160 - 1178.
20 EVAN S, JONATHAN L, TREVOR D. Fully convolutional networks for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3431 - 3440.
21 马树志. 基于深度学习的肝脏 CT 影像分割方法的研究与应用[D]. 长春:吉林大学, 2017.
MA Shuzhi. Research of liver segmentation in CT image based on deep learning [D]. Changchun: Jilin University, 2017. (in Chinese)
22 毋立芳, 贺娇瑜, 简萌, 等. 局部聚类分析的 FCN - CNN 云图分割方法[J]. 软件学报, 2018, 29(4): 1049 - 1059.
WU Lifang, HE Jiaoyu, JIAN Meng, et al. Cloud atlas segmentation method based on FCN and CNN[J]. Journal of Software, 2018, 29(4): 1049 - 1059. (in Chinese)
23 王鹏, 方志军, 赵晓丽, 等. 基于深度学习的人体图像分割算法[J]. 武汉大学学报(理学版), 2017, 63(5): 466 - 470.
WANG Peng, FANG Zhijun, ZHAO Xiaoli, et al. Human segmentation based on deep learning [J]. Journal of Wuhan University (Natural Science Edition), 2017, 63(5): 466 - 470. (in Chinese)
24 KAREN S, ANDREW Z. Very deep convolutional networks for large-scale image recognition [C] // International Conference on Learning Representations (ICLR), 2015: 1 - 14.
25 GARCIA G A, ORTS E S, OPREA S, et al. A review on deep learning techniques applied to semantic segmentation [J/OL]. [2018 - 04 - 22]. https://arxiv.org/abs/1704.06857.