doi:10.6041/j.issn.1000-1298.2020.05.022

基于 BiGRU_MulCNN 的农业问答问句分类技术研究

金　宁[1,2]　赵春江[3,4]　吴华瑞[3,4]　缪祎晟[3,4]　李　思[5]　杨宝祝[3,4]

(1.沈阳农业大学信息与电气工程学院,沈阳 110866;2.沈阳建筑大学研究生院,沈阳 110168;
3.国家农业信息化工程技术研究中心,北京 100097;4.北京农业信息技术研究中心,北京 100097;
5.沈阳建筑大学党委组织部,沈阳 110168)

摘要:“中国农技推广”问答社区每天新增提问数据近万条,对提问的有效分类是实现智能问答的关键技术环节。海量提问数据具有特征稀疏性强、噪声大、规范性差的特点,制约了文本分类效果。为了改善农业问答问句短文本分类性能,提出了 BiGRU_MulCNN 分类模型,运用 TF-IDF 算法拓展文本特征,并加权表示文本词向量,利用双向门控循环单元神经网络获取输入词向量的上下文特征信息,构建多尺度并行卷积神经网络,进行多粒度的特征提取。试验结果表明,基于混合神经网络的短文本分类模型可以优化文本表示和文本特征提取,能够准确地对用户提问进行自动分类,正确率达 95.9%,与其他 9 种文本分类方法相比,分类性能优势明显。

关键词:农业信息分类;自然语言处理;双向门控循环单元神经网络;卷积神经网络

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1000-1298(2020)05-0199-08　　**OSID**:

Classification Technology of Agricultural Questions Based on BiGRU_MulCNN

JIN Ning[1,2]　ZHAO Chunjiang[3,4]　WU Huarui[3,4]　MIAO Yisheng[3,4]　LI Si[5]　YANG Baozhu[3,4]

(1. *School of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang* 110866, *China*
2. *Graduate School, Shenyang Jianzhu University, Shenyang* 110168, *China*
3. *National Engineering Research Center for Information Technology in Agriculture, Beijing* 100097, *China*
4. *Beijing Research Center for Information Technology in Agriculture, Beijing* 100097, *China*
5. *Organization Department, Shenyang Jianzhu University, Shenyang* 110168, *China*)

Abstract: With the rapid development of mobile internet, short text data of APPs has exploded. In the field of agriculture, tens of thousands of questions about agricultural technology have been put forward in agro-technical extension community. Accurate classification is the basis of agricultural intelligent Q&A and the guarantee of precise information service. In order to improve the performance of data classification, a short text classification method based on BiGRU_MulCNN model was proposed to overcome the limitations of the classification process, such as few vocabulary, sparse features, large amount of data, lots of noise and poor normalization. In the model, Jieba word segmentation tools and agricultural dictionary were selected to text segmentation, then TF-IDF algorithm was adopted to expand the text characteristic and weighted word vector according to the text of key vector, and bi-directional gated recurrent unit was applied to catch the context feature information, multi-convolutional neural networks was finally established to gain local multidimensional characteristics of text. Batch-normalization, Dropout, Global Average Pooling and Global Max Pooling were involved to solve over-fitting problem. The results showed that the model could classify questions accurately, with an accuracy of 95.9%. Compared with other models, such as CNN model, RNN model and CNN/RNN combinatorial model, BiGRU_MulCNN had obvious advantages in classification performance in intelligent agro-technical information service.

Key words: classification of agriculture information; natural language processing; bi-directional gated recurrent unit; convolutional neural network

收稿日期:2019-08-20　修回日期:2019-11-26

基金项目:国家自然科学基金项目(61871041、61571051)和北京市自然科学基金项目(4172024、4172026)

作者简介:金宁(1989—),男,博士生,沈阳建筑大学助理研究员,主要从事农业智能系统研究,E-mail:jinning21@126.com

通信作者:赵春江(1964—),男,研究员,中国工程院院士,主要从事农业人工智能与智能系统研究,E-mail:zhaocj@nercita.org.cn

0 引言

随着移动互联网产业的高速发展,各类移动应用程序产生的评论信息、微信朋友圈、问答社区用户提问等短文本数据呈爆发式增长[1]。在农业领域,“中国农技推广 App”作为农业信息服务方面的移动应用程序,为农业技术人员及农户搭建了学习交流平台,帮助农户实时获得在线农业技术指导。但大多数农户不会选择提问分类,部分已选择的也存在分类不准确的问题,从而影响了农业技术指导的高效性、精确性。“中国农技推广”每天增衍提问数量近万条,人工筛选将消耗大量的人力、物力,且无法高效、准确实现分类。因此,利用计算机技术解决农户提问的自动分类是“中国农技推广”当前亟需解决的问题。农业问答问句的自动分类是实现农业智能问答的关键技术环节,是自然语言处理和农业大数据智能研究领域的热点研究方向。

目前,深度学习方法[2-5]和机器学习方法[6-8]在解决文本分类问题上均取得了一定成果。在深度学习方法中,KIM[9]将文本当作固定长度的图像,运用卷积神经网络(Convolutional neural networks, CNN)有效解决了文本分类问题。随后,研究人员以此为基础,不断优化文本分类模型,捕获高层次的文本特征[10-11]。由于 CNN 模型未考虑文本的语序,因此无法获得文本上下文信息,制约了文本分类效果。相比于 CNN 模型,循环神经网络(Recurrent neural network, RNN)模型可对前后信息进行记忆,并应用于当前的计算,更适合处理序列化的文本数据, MIKOLOV 等[12]运用 RNN 模型实现了文本分类。但 RNN 长期依赖学习特征,容易出现梯度弥散的问题,为此,研究人员提出了长短期记忆神经网络(Long/short term memory, LSTM)和门控循环单元神经网络(Gated recurrent units, GRU)等优化模型,并应用于文本分类问题,取得了较好的分类效果。RNN 模型存在计算复杂、内存占用大、训练耗时长、对局部关键信息提取不敏感等问题。为更好地提取文本关键信息,注意力机制(Attention)[13]被广泛应用于文本分类问题[14-15],其通过模仿人脑的注意力分配机制,计算不同词向量的权重,使关键词语的权重更高,从而获得高质量的文本特征。在机器学习方法中,K 最近邻算法[16]、朴素贝叶斯模型[17]、隐马尔科夫模型[18]等方法广泛应用于文本分类,但存在严重的数据稀疏问题,影响了分类效果。在农业领域,由于受农业大数据源问题的限制,相关研究仍处于起步阶段。魏芳芳等[19]运用支持向量机算法,周云成等[20]运用朴素贝叶斯算法,实现了机器学习算法对中文农业长文本的自动分类。由于机器学习方法需要人工提取特征,使其特征工程往往仅适用于特定数据集,不具备深度学习方法的适应性和易迁移性。此外,赵明等[21]针对番茄病虫害问答系统问句分类问题,提出了基于双向门控循环单元神经网络(Bi-directional gated recurrent unit, BiGRU)的短文本分类模型,分类准确率明显提升;梁敬东等[22]利用 LSTM 算法计算问句相似度,提高了问答系统回答的准确性;许童羽等[23]提出一种基于注意力机制优化的序列到序列(Sequence to sequence, Seq2Seq)问答模型,提高了水稻病虫害问答的准确率;张明岳等[24]利用 CNN 提取文本特征,用于判断问句是否有效,识别准确率明显提升。上述研究为深度学习方法在农业领域的文本分类提供了可行性依据和参考,但在文本特征提取方面仍存在不足,特征提取方法较为单一,未能有效解决短文本的特征不足问题,并且相关模型均在特定的农业领域应用,未在涉及多类别的农业问答数据集中进行验证。

针对农业问答问句短文本词汇量少、特征稀疏性强、数据量大、噪声大、规范性差的特点,本文对短文本特征词汇进行拓展,根据词汇重要程度加权表示词向量,利用 BiGRU 和 CNN 提取文本特征,进一步优化和改进神经网络模型结构及参数,构建一种基于混合神经网络的短文本分类模型,以实现农业问答问句在多个类别上的精准自动分类。

1 BiGRU_MulCNN 文本分类模型构建

本文提出的 BiGRU_MulCNN 模型如图 1 所示。该模型主要由文本预处理层、双向门控循环单元层(BiGRU)和多尺度卷积神经网络层(MulCNN)3 部分组成。与传统深度学习分类模型相比,本文所提分类模型增加了对文本的加权预处理,使用 TF-IDF 算法扩充文本特征词语,根据词语的重要程度计算加权词向量;采取多种方法提取文本特征,利用 BiGRU 获取词语的上下文信息,构建多尺度并行 CNN 以便提取文本不同粒度的局部特征。

1.1 文本预处理

由于计算机无法将中文文本直接作为分类模型的输入进行分类计算,因此需要先将中文文本转换成数字向量。为了尽可能保留文本特征及语义信息的完整性、全面性,本文首先对提问文本进行去噪、分词等预处理操作,然后运用 Word2vec[25]方法将分词结果转换为词向量。本文提出的文本预处理流程如图 2 所示。

1.1.1 文本分词

本文采用 Python 的 Jieba 分词库对文本进行分

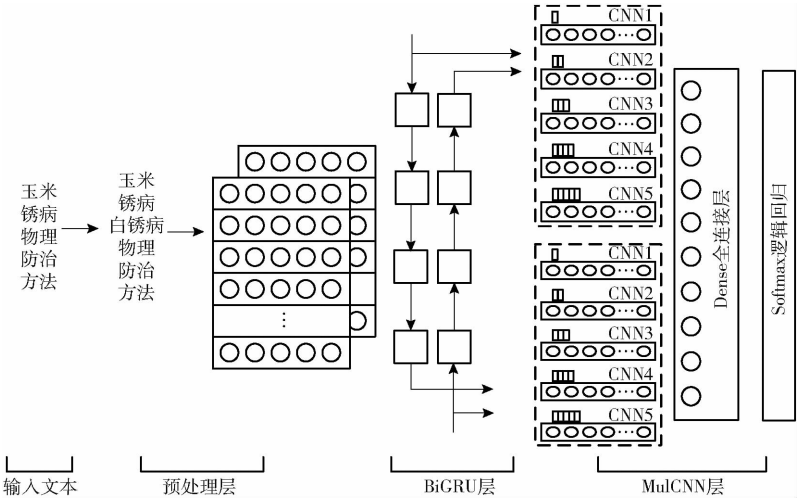

图 1　BiGRU_MulCNN 模型结构图

Fig. 1　Schematic of BiGRU_MulCNN

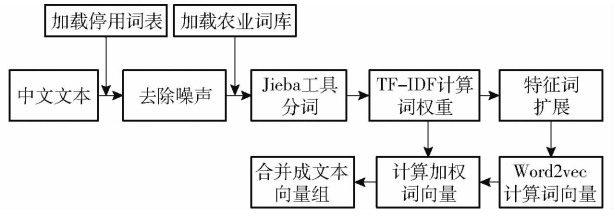

图 2　文本预处理流程图

Fig. 2　Flow chart of data preprocessing

词。由于中文的分词结果受语义和语境影响较大，为提高分词的准确度，在分词前加载了停用词表，去除文本中的停用词、特殊字符及空格等不利于特征提取的噪声，减少文本的冗余信息[26]。针对农业问答数据集专业词汇多的特点，本文加载了搜狗农业词汇大全作为分词字典[27]代替基础分词库，提高对农业专业词汇的识别度。

1.1.2　特征词扩展

扩展短文本的特征是提高分类正确率的有效方法[28]。问句中每个词语的重要程度均不相同，重要程度高的词语更能体现提问的语意，更具有代表性。本文采用 TF－IDF 方法计算每个词语的重要程度，提取问句中最具有代表性的特征词。TF－IDF 方法可保留文本中具有代表性的低频词语，去除区分度低的高频词，词频(TF)表示词语在全部词语中出现的频率，计算公式为

$$f_{i,j}=\frac{n_{i,j}}{\sum_{m} n_{m,j}} \tag{1}$$

式中　$f_{i,j}$——词 t_i 的词频

$n_{i,j}$——词 t_i 在文本中出现的次数

m——文本包含的单词数

$n_{m,j}$——词 t_m 在文本中出现的次数

逆文档频率 IDF 表示词语的普遍程度，计算公式为

$$q_i=\lg\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{2}$$

式中　q_i——逆文档频率

D——文本总量

d_j——包含词 t_i 的文本

j——包含词 t_i 的文本数量

TF－IDF 值为 $f_{i,j}$ 与 q_i 的乘积，计算公式为

$$s_{i,j}=f_{i,j}q_i \tag{3}$$

式中　$s_{i,j}$——词 t_i 的 TF－IDF 值

将每个问句中 TF－IDF 值最高的词语作为问句的特征词。计算其他词语与该特征词的相似度，选择相似度高于 80% 且排序前 5 的词语对文本进行特征词扩充。

1.1.3　加权词向量表示

Word2vec 是近年来较为流行的中文文本分布式表示方法[29]。Word2vec 可根据输入的目标词语，预测上下文信息，并将意思相近的词映射到向量空间中相近的位置，有效解决了 One－Hot 方法词向量相互孤立和维度高的问题。本文采用 Word2vec 方法的 Skip－gram 模型训练分词结果，将中文词语转换为低维、连续的词向量。

为进一步突出不同词语对问句含义的贡献程度，本文将词语的 TF－IDF 值与 Word2vec 词向量的乘积作为该词语的加权词向量。

1.2　文本表示

本文先将问句中包含词语的加权词向量连接起来，组成加权文本向量组，将其作为 BiGRU 模型的输入。为充分考虑中文文本语序对语义的影响，本文利用双向门控循环单元神经网络挖掘当前词语的上下文信息，获得表达更为精确、特征更为丰富的文本向量，最后将 BiGRU 模型的输出与原加权文本向量组连接，组成新的文本向量。

1.2.1 加权词向量文本

获得每个词的加权词向量后,将文本中的每个词替换成其对应的词向量,组成加权文本向量组。由于问句的长短不一,需统一问句长度后,方可输入到神经网络模型中训练。根据对文本的统计,99.9%的问句长度均少于 100 个词,因此将文本问句的长度设置为 100,其余提问长度不足的,填充 0 补齐文本向量,长度超过 100 的只取前 100 个词。门控循环单元神经网络结构图如图 3 所示。

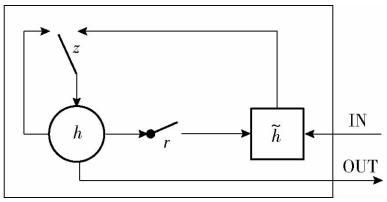

图 3 门控循环单元神经网络结构图

Fig. 3 Schematic of GRU

1.2.2 BiGRU 模型文本

GRU[30]是一种特殊的循环神经网络,能够有效解决循环神经网络中无法长期记忆和反向传播的梯度问题。与 LSTM 相比,GRU 具有参数少、结构简单、便于计算、收敛性强的特点,其具体结构如图 3 所示。

GRU 结构中包含 2 种状态和 2 个控制门,分别是隐含状态 h、候选状态 $\tilde{h}$、重置门 r 和更新门 z,其中更新门用于控制前一时刻的状态信息传入到当前状态中的程度,重置门用于控制忽略前一时刻状态信息的程度。在 t 时刻,$\tilde{h}_t$ 的计算依赖于输入词向量 $\boldsymbol{x}_t$ 和 h_{t-1},r_t 作用于 h_{t-1},并根据 h_{t-1} 的重要程度控制过去隐含状态保留程度。r_t 越大,表示 h_{t-1} 对 $\tilde{h}_t$ 的影响程度越大。GRU 参数计算公式为

$$r_t = \sigma_g w_r(\boldsymbol{x}_t, h_{t-1}) \tag{4}$$

$$z_t = \sigma_g w_z(\boldsymbol{x}_t, h_{t-1}) \tag{5}$$

$$\tilde{h}_t = \tanh(\boldsymbol{w}(\boldsymbol{x}_t, r_t \odot h_{t-1})) \tag{6}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{7}$$

式中 w_r——重置门权重 $\boldsymbol{x}_t$——输入词向量

w_z——更新门权重

h_t——隐含层状态

r_t——重置门 z_t——更新门

σ_g——Sigmoid 函数

$\boldsymbol{w}$——权重矩阵

$\odot$——对应元素相乘符号

GRU 神经网络是从前向后单向输出的。这与中文语意理解方式略有不同,中文语意与当前文字的上下文均有关系。在文本分类任务中,如果当前时刻的输出能与前后时刻的状态都产生联系,会更有利于文本深层次特征的提取,突出文本关键信息。根据中文语意理解的特点,本文利用 BiGRU 模型提取问句的特征向量。BiGRU 模型是由两个单向且方向相反的 GRU 组成的神经网络模型,其输出由两个不同方向的 GRU 的状态共同决定。文本在 t 时刻输入的第 i 个句子的第 j 个单词的词向量为 $\boldsymbol{c}_{tij}$,其隐含层状态 h_t 由前向隐含层状态 h_{ft-1} 和反向隐含层状态 h_{rt-1} 加权得到,计算过程为

$$h_{ft} = \mathrm{GRU}(\boldsymbol{c}_{tij}, h_{ft-1}) \tag{8}$$

$$h_{rt} = \mathrm{GRU}(\boldsymbol{c}_{tij}, h_{rt-1}) \tag{9}$$

$$h_t = \boldsymbol{y}_t h_{ft} + \boldsymbol{v}_t h_{rt} + b_t \tag{10}$$

式中 GRU(·)——词向量的非线性变换函数

$\boldsymbol{y}_t$——前向权重矩阵

$\boldsymbol{v}_t$——反向权重矩阵

b_t——偏置

1.3 MulCNN 模型构建

在 BiGRU 模型获得词语上下文信息后,构建了 MulCNN 模型,进一步提取文本高维度、多尺度的局部特征。MulCNN 模型由多个一维卷积层、池化层、全连接层和分类层组成。

1.3.1 卷积层

卷积层的作用是在设定的窗口范围提取局部特征,利用卷积核对输入向量进行卷积计算,获得特征输出。在一维卷积神经网络中,卷积核长度为词向量的维度,高度为设定窗口的大小,卷积计算公式为

$$c_j = f(\boldsymbol{x}_j k + b) \tag{11}$$

式中 c_j——窗口特征值

f——激活函数 $\boldsymbol{x}_j$——词向量

k——卷积核 b——偏置

针对短文本语义依赖距离短的特点[31],为了能够提取文本的多粒度局部特征,本文设置了宽度不同、数量不同的多个卷积核窗口的卷积神经网络。将不同粒度的特征值合并,作为卷积层计算的特征值。

1.3.2 池化层

由于在卷积层选择了多个不同窗口宽度、不同数量的卷积核,使得卷积计算后生成的特征图维度不一致,因此本文在模型中增加了池化层。池化层将卷积层提取的文本局部特征进一步整合,在缩减特征图尺寸、提高计算速度的同时,使特征值获得了全局信息,提高了所提取特征的鲁棒性,控制了过拟合问题发生。本文利用全局平均池化和全局最大池化方法进行池化操作,即抽取每个特征图的最大值和平均值,将两者拼接后作为该特征图的特征值。

1.3.3 全连接层

全连接层进一步对特征值进行抽象,将池化层

的全部输出作为输入，其中每一个神经单元都与池化层的每一个单元对接，并通过激活函数 ReLu 将池化层向量转换成长向量，将文本从特征空间映射到标记空间。

1.3.4　分类层

使用 Softmax 函数作为特征分类器。Softmax 函数对全连接层的输出进行归一操作，映射到(0,1)区间内，得到每类特征输出的估算值。

2　试验与结果分析

2.1　试验数据

从“中国农技推广”农技问答模块 2019 年不同月份的提问数据中随机提取 20 000 条作为试验数据，提问类别具体分布情况见表 1。由表 1 可知，试验数据涉及类别多，覆盖了病虫草害、栽培管理、养殖管理等 12 个类别，并且数据分布不平衡，病虫草害、栽培管理等类别数据量达几千条，而屠宰加工等类别数据量仅有几十条，数据量相差悬殊，增加了文本分类的难度。

从每个类别的问句中随机选择 10% 作为测试数据集，共 2 000 条。在剩余数据中每个类别选择 90% 的数据作为训练数据集，共 16 200 条；10% 的数据作为验证数据集，共 1 800 条，用于验证模型训练及优化情况。测试数据集、训练数据集和验证数据集均无重复交叉，因此测试数据集的试验结果可作为模型分类效果的评价指标。

表 1　问题类别分布

Tab.1　Distribution of question category

问题类别	病虫草害	市场销售	动物疫病	栽培管理	养殖管理	土壤肥料	饲料营养	采收加工	农业机械	贮运保鲜	屠宰加工	其他
数量/条	6 702	443	2 044	4 284	991	840	128	137	309	127	28	3 967

2.2　参数设置

使用 128 维词向量表示中文词汇，设置问句最大长度为 100。BiGRU 层设定 GRU 输出特征维度为 128 维，并选择 concat 模式连接 GRU 的前向和后向输出。

由 1.3 节可知，MulCNN 模型在同一窗口下包含多组卷积核个数不同的卷积神经网络。试验中，相同窗口下设置了 2 组卷积神经网络，不同数量的卷积核得到的试验结果见表 2。当卷积核尺寸为(96,160)时，分类效果最佳。

表 2　MulCNN 模型卷积核的确定

Tab.2　Determination of kernel size in MulCNN

卷积核尺寸	(64,96)	(64,128)	(64,160)	(96,128)	(96,160)	(128,128)
正确率/%	95.50	95.20	94.45	95.60	95.65	94.95

设计多个卷积窗口尺寸不同的卷积层，用于提取不同粒度的文本特征。具体卷积窗口尺寸设置情况及试验结果如表 3 所示。当卷积窗口数为 5，窗口宽度为 1、2、3、4、5 时取得了最好的分类效果。

表 3　MulCNN 模型卷积窗口尺寸的确定

Tab.3　Determination of filters in MulCNN

窗口数	窗口宽度	正确率/%
3	2、3、4	95.65
4	1、2、3、4	95.50
4	2、3、4、5	95.80
5	1、2、3、4、5	95.90
5	2、3、4、5、6	95.50

为防止过拟合，对 BiGRU 和 MulCNN 均进行批规范化处理，全连接层单元丢弃比例设定为 0.5，训练过程中通过降低神经网络的学习率来提高性能，每隔 10 次训练 1 次学习率减小到原来的 1/10。

2.3　对比模型

将 BiGRU_MulCNN 与 9 种近年来在文本分类领域和农业领域常用的分类模型进行比较，9 种分类模型可总结为 CNN 分类模型、RNN 分类模型和混合神经网络分类模型 3 类。

CNN 分类模型：TextCNN 模型是将 CNN 首次应用于文本分类的模型；DCNN[32] 模型利用 K 最大池化的动态 CNN 进行文本分类；DPCNN[33] 模型利用深层 CNN 进行文本分类；Agro－CNN 模型[24] 是针对农业问答有效性的识别模型。

RNN 分类模型：TextRNN[34] 模型利用标准 LSTM 进行文本分类；AttBiRNN[35] 模型利用 BLSTM 并引入注意力机制进行文本分类；N－BiGRU 模型[21] 是针对番茄病虫害问答系统的多层 BiGRU 分类模型。

混合神经网络分类模型：RCNN[36] 模型利用前向和后向 RNN 结合 CNN 进行文本分类；C－LSTM[37] 模型利用 CNN 获得高维度词表示，结合 LSTM 进行文本分类。图 4 为不同模型下文本分类正确率的对比。

2.4　结果分析

图 4 展示了 10 种试验模型在 Word2vec 文本及 TF－IDF 加权文本表示下的文本分类正确率。正确率是对全部数据集分类结果准确性的判断，一般用于衡量模型的整体分类效果。如图 4 所示，针对农

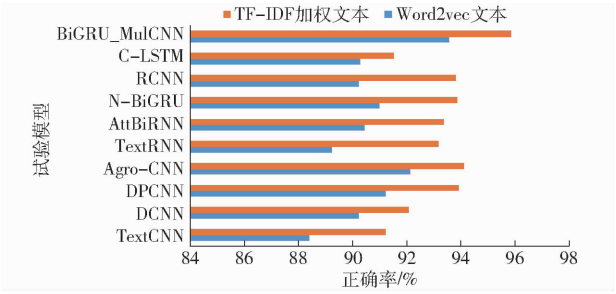

图 4 不同模型下文本分类正确率对比

Fig. 4 Comparison of accuracy in different models

业问答问句短文本数据集,本文提出的 TF - IDF 加权文本表示方式在 10 种试验模型的分类正确率均大幅超过 Word2vec 文本表示方式,特别对于 RNN 分类模型的正确率提升明显。本文提出的 BiGRU_MulCNN 模型在 Word2vec 文本表示方式和 TF - IDF 加权文本表示方式下均取得了最优的结果,正确率分别达到了 93.60% 和 95.90%,相比于其他 9 种对比模型优势显著。在 TF - IDF 加权文本表示方式下,CNN 分类模型中 Agro - CNN 正确率最高,达到 94.15%;RNN 分类模型中 N - BiGRU 正确率最高,达到 93.90%;混合神经网络模型中 RCNN 正确率最高,达到 93.85%。

图 5 展示了在 TF - IDF 加权文本表示下,各个类别分类模型中正确率最高的 Agro - CNN、N - BiGRU、RCNN、BiGRU_MulCNN 模型对 12 个问题类别分类的 F1 值。F1 值表示分类精确率和召回率的调和平均数,常用于衡量模型分类性能。如图 5 所示,BiGRU_MulCNN 模型的 F1 值在病虫草害、市场营销、动物疫病等 9 个类别中均为最高,整体分类效果明显优于其他模型。在病虫草害、栽培管理等试验数据量充足的数据集中,本文模型的 F1 值略高于其他模型;在动物疫病、养殖管理、农业机械等数据量较少的数据集中,本文模型的 F1 值明显高于其他模型,说明 BiGRU_MulCNN 模型在数据量不充足的情况下,仍然能够有效提取短文本的特征进行分类。但是在饲料营养、屠宰加工等试验数据集过少的情况下,4 种试验模型表现均不稳定,说明深度学习分类模型需要大量数据集支撑,数据量过小会影响分类效果。

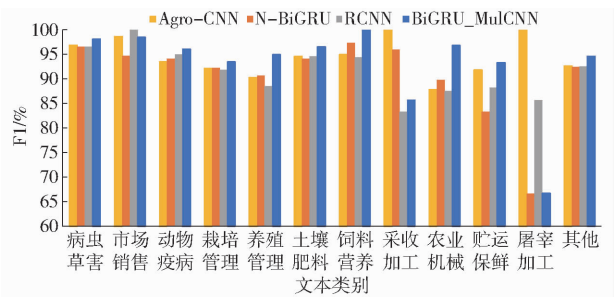

图 5 4 种试验模型对于不同问题类别分类的 F1 值对比

Fig. 5 Comparison of F1 values of four models for different question categories

在 12 个问题类别中,栽培管理类别虽然试验数据量充足,但分类效果远不如数据量较少的市场营销、动物疫病等类别。通过分析试验文本可知,栽培管理类别涵盖了多种复杂的农业生产操作,覆盖面过大,导致了该类别的特征不明显,影响了分类效果。表 4 统计了 4 种试验模型在栽培管理类别的精确率、召回率和 F1。由表可知,BiGRU_MulCNN 模型的精确率、召回率和 F1 均取得了较好的结果,其中精确率和 F1 远远高于其他模型,说明了该模型具有较强的鲁棒性。

表 4 4 种试验模型在栽培管理类别的比较

Tab. 4 Comparison of four models in cultivation management categories %

试验模型	精确率	召回率	F1
Agro - CNN	92.07	92.29	92.18
N - BiGRU	90.93	93.69	92.29
RCNN	92.27	92.06	92.16
BiGRU_MulCNN	94.21	92.92	93.56

如图 6 所示,试验数据集的规模直接影响了模型的分类正确率。随着数据量的增加,各分类模型的正确率均随之增加,其中 N - BiGRU 和 BiGRU_MulCNN 在数据量较小的情况下分类效果较好,BiGRU_MulCNN 模型在大数据集上分类效果优势明显。

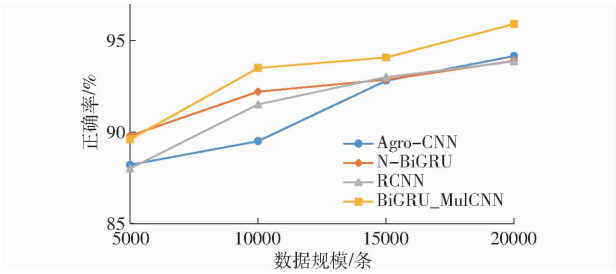

图 6 不同数据规模的模型分类正确率

Fig. 6 Classification accuracy of models for different datasets

如表 5 所示,4 种试验模型对 2 000 条测试数据的响应时间达到了快速反馈问题分类的要求。其中 RCNN 模型由于其结构简单,模型层数较少,在训练时间和测试时间上的优势明显,但其正确率明显低于 BiGRU_MulCNN 模型;以 CNN 为基础的 Agro - CNN 模型的训练时间较短,以 RNN 为基础的 N - BiGRU 模型及基于混合神经网络的 BiGRU_MulCNN 模型的训练时间较长。由于分类模型的训练可以离线运行,在分类结果反馈时间基本相同的情况下,分类模型更关注分类正确率的提升。

表 5　4 种试验模型的离线训练时间和测试响应时间

Tab. 5　Offline training time and test time comparison of four models　　s

试验模型	测试响应时间	离线训练时间
Agro－CNN	6	3 425
N－BiGRU	7	4 914
RCNN	3	675
BiGRU_MulCNN	7	6 620

3　结论

(1)提出的 BiGRU_MulCNN 模型满足实际应用需求,可有效解决农业问答问句在多个类别上的自动分类问题,对测试集的正确率达到 95.9%,大幅提高了分类正确率。在数据量不足、数据特征不明显的数据集上仍取得了较好的分类效果,切实解决了传统人工分类耗时、耗力的问题,实现了对农业问答问句的智能分类。

(2)对短文本进行特征词扩充,并根据词语重要性对文本词向量进行加权表示,可明显提高分类的正确率,有效解决了短文本特征不足的问题。

参　考　文　献

[1] 中国互联网络信息中心第 43 次中国互联网络发展状况统计报告[R/OL]. [2019－08－07]. http://www.cac.gov.cn/wxb_pdf/0228043.pdf.

[2] BENGONG Y, LINBIN Z. Chinese short text classification base on CP－CNN [J]. Application Research of Computers, 2018, 4: 1001－1004.

[3] ZHAO W, YE J, YANG M, et al. Investigating capsule networks with dynamic routing for text classification[J]. arXiv Preprint:1804.00538, 2018.

[4] PAPPAS N, POPESCU-BELIS A. Multilingual hierarchical attention networks for document classification[J]. arXiv Preprint: 1707.00896, 2017.

[5] 郑诚,洪彤彤,薛满意.用于短文本分类的 BLSTM_MLPCNN 模型[J].计算机科学,2019,46(6):206－211.
ZHENG Cheng, HONG Tongtong, XUE Manyi. BLSTM_MLPCNN model for short text classification[J]. Computer Science, 2019,46(6):206－211. (in Chinese)

[6] 靳一凡,傅颖勋,马礼.基于频繁项特征扩展的短文本分类方法[J].计算机科学,2019,46(增刊):478－481.
JIN Yifan, FU Yingxun, MA Li. Method of short text classification based on frequent item feature extension[J]. Computer Science, 2019,46(Supp.):478－481. (in Chinese)

[7] 古丽娜孜·艾力木江,乎西旦·居马洪,孙铁利,等.基于支持向量的最近邻文本分类方法[J].智能系统学报,2018, 13(5):799－807.
GULNAZ Alimjan, HURXIDA Jumahun, SUN Tieli, et al. The nearest neighbor text classification method based on support vector[J]. CAAI Transactions on Intelligent Systems, 2018,13(5):799－807. (in Chinese)

[8] LIU P, ZHAO H H, TENG J Y, et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark [J]. Journal of Central South University, 2019, 26(1):1－12.

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

[10] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C]// Advances in Neural Information Processing Systems, 2015: 649－657.

[11] DOS S C, GATTI M. Deep convolutional neural networks for sentiment analysis of short texts[C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers,2014: 69－78.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems, 2013: 3111－3119.

[13] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C]// Advances in Neural Information Processing Systems, 2014: 2204－2212.

[14] ROZENTAL A, FLEISCHER D. Amobee at semeval-2018 task 1: GRU neural network with a CNN attention mechanism for sentiment classification[J]. arXiv Preprint:1804.04380, 2018.

[15] CHEN P, SUN Z, BING L, et al. Recurrent attention network on memory for aspect sentiment analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,2017: 452－461.

[16] De VRIES A P, MAMOULIS N, NES N, et al. Efficient k-NN search on vertically decomposed data[C]// Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. ACM, 2002: 322－333.

[17] CHEN Z, SHI G, WANG X. Text classification based on Naive Bayes algorithm with feature selection[J]. International Information Institute (Tokyo). Information, 2012, 15(10): 4255.

[18] VIEIRA A S, BORRAJO L, IGLESIAS E L. Improving the text classification using clustering and a novel HMM to reduce the dimensionality[J]. Computer Methods and Programs in Biomedicine, 2016, 136: 119－130.

[19] 魏芳芳,段青玲,肖晓琰,等.基于支持向量机的中文农业文本分类技术研究[J/OL].农业机械学报,2015,46(增刊): 174－179.
WEI Fangfang, DUAN Qingling, XIAO Xiaoyan, et al. Classification technique of Chinese agricultural text information based on SVM[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015,46(Supp.):174－179. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=2015S029&journal_id=jcsam DOI: 10.6041/j.issn.1000-1298.2015.S0.029. (in Chinese)

[20] 周云成,许童羽,邓寒冰.基于 NB 和 CHI 值的农业文本分类方法[J].江苏农业科学,2018,46(17):219－223.

[21] 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.
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. 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. (in Chinese)

[22] 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)

[23] 许童羽,赵冬雪,周云成,等. 基于 word2vec 和 Attention - Seq2Seq 的水稻病虫害智能问答方法研究[J]. 沈阳农业大学学报,2019,50(3):378 - 384.
XU Tongyu, ZHAO Dongxue, ZHOU Yuncheng, et al. Research on method of intelligent Q & A for rice pests and diseases based on word2vec and Attention - Seq2Seq[J]. Journal of Shenyang Agricultural University, 2019, 50(3): 378 - 384. (in Chinese)

[24] 张明岳,吴华瑞,朱华吉. 基于卷积模型的农业问答语性特征抽取分析[J/OL]. 农业机械学报,2018,49(12):203 - 210.
ZHANG Mingyue, WU Huarui, ZHU Huaji. Analysis of extraction of semantic feature in agricultural question and answer based on convolutional model[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(12): 203 - 210. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20181226&flag = 1. DOI: 10.6041/j.issn.1000-1298.2018.12.026. (in Chinese)

[25] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint: 1301.3781, 2013.

[26] 官琴,邓三鸿,王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现,2017,1(3):72 - 80.
GUAN Qin, DENG Sanhong, WANG Hao. Chinese stopwords for text clustering: a comparative study[J]. Data Analysis and Knowledge Discovery, 2017, 1(3): 72 - 80. (in Chinese)

[27] 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)

[28] WANG X, CHEN R, JIA Y, et al. Short text classification using wikipedia concept based document representation[C]// 2013 International Conference on Information Technology and Applications. IEEE, 2013: 471 - 474.

[29] CHENGZHANG X, DAN L. Chinese text summarization algorithm based on Word2vec[C]// Journal of Physics: Conference Series. IOP Publishing, 2018, 976(1): 012006.

[30] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv Preprint:1406.1078, 2014.

[31] MA M, HUANG L, XIANG B, et al. Dependency-based convolutional neural networks for sentence embedding[J]. arXiv Preprint:1507.01839, 2015.

[32] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. arXiv Preprint:1404.2188, 2014.

[33] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 562 - 570.

[34] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[J]. arXiv Preprint:1605.05101, 2016.

[35] RAFFEL C, ELLIS D P W. Feed-forward networks with attention can solve some long-term memory problems[J]. arXiv Preprint:1512.08756, 2015.

[36] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]// Twenty-ninth AAAI Conference on Artificial Intelligence, 2015.

[37] ZHOU C, SUN C, LIU Z, et al. A C - LSTM neural network for text classification[J]. Computer Science, 2015, 1(4): 39 - 44.