doi:10.6041/j.issn.1000-1298.2021.03.029

基于深度卷积神经网络的水稻知识文本分类方法

冯　帅[1]　许童羽[1,2]　周云成[1,2]　赵冬雪[1]　金　宁[1]　王郝日钦[1]
(1.沈阳农业大学信息与电气工程学院,沈阳 110161;2.沈阳农业大学辽宁省农业信息化工程技术中心,沈阳 110161)

摘要:为解决文本特征提取不准确和因网络层次加深而导致模型分类性能变差等问题,提出基于深度卷积神经网络的水稻知识文本分类方法。针对水稻知识文本的特点,采用 Word2Vec 方法进行文本向量化处理,并与 One-Hot、TF-IDF 和 Hashing 方法进行对比分析,得出 Word2Vec 方法具有较高的分类精度,正确率为 86.44%,能够有效解决文本向量表示稀疏和信息不完整等问题。通过调整残差网络(Residual network,ResNet)结构,分析残差模块结构和网络层次对分类网络的影响,构建了 9 种分类网络结构,测试结果表明,具有 4 层残差模块结构的网络具有较好的特征提取精度,Top-1 准确率为 99.79%。采用优选出的 4 层残差模块结构作为基本结构,使用胶囊网络(Capsule network,CapsNet)替代其池化层,设计了水稻知识文本分类模型。与 FastText、BiLSTM、Atten-BiGRU、RCNN、DPCNN 和 TextCNN 等 6 种文本分类模型的对比分析表明,本文设计的文本分类模型能够较好地对不同样本量和不同复杂程度的水稻知识文本进行精准分类,模型的精准率、召回率和 F1 值分别不小于 95.17%、95.83% 和 95.50%,正确率为 98.62%。本文模型能够实现准确、高效的水稻知识文本分类,满足实际应用需求。

关键词:水稻知识文本;文本分类;深度卷积神经网络;向量化处理;特征提取;分类模型
中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1000-1298(2021)03-0257-08　　**OSID**:

Rice Knowledge Text Classification Based on Deep Convolution Neural Network

FENG Shuai[1]　XU Tongyu[1,2]　ZHOU Yuncheng[1,2]　ZHAO Dongxue[1]　JIN Ning[1]　WANG Haoriqin[1]
(1. *College of Information and Electrical Engineering*, *Shenyang Agricultural University*, *Shenyang* 110161, *China*
2. *Liaoning Agricultural Information Technology Center*, *Shenyang Agricultural University*, *Shenyang* 110161, *China*)

Abstract: The data of weeds, pests, diseases and cultivation management of rice extracted from agricultural text data is a typical text classification problem, which is fundamental to key text information extraction, text data mining and agricultural intelligent question and answer. The classification of Chinese texts, especially agricultural texts, is characterized by poor data redundancy, sparsity and normativity. While the deep learning technology can automatically extract the key features of the text, and the built model has strong adaptability and mobility. For that reason, in order to solve the problem of classification performance of the model deteriorates caused by inaccurate text feature extraction and deepened network hierarchy, a text classification method of rice knowledge oriented Q&A system was proposed. The Python of scrapy was adopted to obtain Chinese text data on rice pests, grass pests, cultivation and management, such as the experts online system of Hownet and the planting question and answer website, as training and test samples. Jieba segmentation method was applied to rice knowledge text for word segmentation to remove useless symbols and stop words in the text. Meanwhile, the results of Chinese segmentation were greatly influenced by the segmentation lexicon. In order to improve the precision of word segmentation of rice knowledge text and reduce the situation of misclassification, omission and misclassification, a rice-related corpus was constructed on the basis of sogou agricultural corpus, which further expanded the basic Jieba word segmentation database and improved the identification degree of specialized words such as rice diseases, insect pests, grass and drugs, cultivation and management. At the same time, Word2Vec method was used to vectorize text data, and it was compared with One-Hot, TF-IDF and Hashing

收稿日期:2020-06-13　修回日期:2020-08-09
基金项目:国家重点研发计划项目(2018YFD0300309)
作者简介:冯帅(1992—),男,博士生,主要从事自然语言处理研究,E-mail:int_crazy@163.com
通信作者:许童羽(1967—),男,教授,博士生导师,主要从事农业信息化研究,E-mail:yatongmu@163.com

methods, and it was concluded that Word2Vec method can effectively solve the text vector typical problems such as sparsity and incomplete information. Based on the fundamental structure of ResNet, nine kinds of rice knowledge text classification models were constructed by means of the change and design of its residual module and network hierarchy. The test results indicated that a network with 4-layer residual module structure had good feature extraction accuracy, and the Top－1 accuracy was 99.79%. In the convolutional neural network, the pooling layer was used for the under-sampling operation, which would lose certain text phrase relative position characteristics in the pooling process, thus affecting the classification accuracy of the model, therefore, the optimized 4-layer residual module structure was taken as the basic structure, and the CapsNet was used to replace the pooling layer, and a rice knowledge text classification model, referred to as RIC－Net, was designed. Through comparative analysis of six text classification models, including FastText, BiLSTM, Atten－BiGRU, RCNN, DPCNN and TextCNN, it was concluded that the text classification model designed was able to precisely classify rice knowledge texts with different sample sizes and different levels of complexity, which enabled the accuracy rate, recall rate and F1 value of the model to be no less than 95.17%, 95.83% and 95.50%, respectively, and the accuracy rate was as high as 98.62%. The model can realize accurate and efficient classification of rice knowledge text, meeting practical application requirements.

Key words: rice knowledge text; text classification; deep convolution neural network; vectorization; feature extraction; classification model

0 引言

在农业智能问答系统中,由于大量的知识文本数据具有稀疏性强、噪声大及类别繁杂等特点,导致所构建的问答系统的准确率较低。因此,利用计算机技术提取准确的文本特征、实现知识文本的自动分类是构建农业智能问答系统的关键技术环节。从农业文本数据中提取出水稻的草害药害、病虫害以及栽培管理等数据是典型的文本分类问题。目前 K 最近邻(K-nearest neighbor, KNN)[1]、朴素贝叶斯(Naive Bayesian, NB)[2]以及支持向量机(Support vector machine, SVM)[3]等机器学习方法是进行文本分类的常用方法。文献[4]采用朴素贝叶斯算法实现了对农业文本的自动分类,识别率较高,但该方法缺乏较好的特征提取能力。文献[5]采用粒子群算法优化 KNN 算法的特征权重,构建 PSOKNN 文本分类模型。文献[6－7]根据农业文本数据特征构建农业行业词库,并通过特征词筛选和权重计算构建一种基于线性支持向量机的中文农业文本分类模型,该方法并未考虑数据集线性不可分的情况,存在一定的局限性。农业文本具有数据冗余性、稀疏性和规范性差等特征,采用传统的机器学习方法对大数据量的农业文本进行分类难度较大,且适应性较差,特征工程复杂。

随着计算机技术的迅猛发展,深度卷积神经网络(Deep convolutional neural networks, DCNN)[8]、循环神经网络(Recurrent neural network, RNN)[9]和胶囊网络(CapsNet)[10]等深度学习技术逐渐成为主流分类方法[11－16]。该技术能够自动实现图像和文本关键特征的提取,无需复杂的特征工程,与分类过程结合,所构建的模型具有较强的适应性和迁移性[17－18]。目前,国内外学者采用深度学习技术在英文和中文文本分类上进行了大量研究[19－23]。相比传统的文本分类方法,深度学习技术在文本分类中具有更好的分类效果。但在文本分类过程中仍存在文本特征提取不准确的问题,如忽略了文本的位置特征等。此外,也未见对网络层次加深所导致的文本分类性能变差的原因进行研究与分析的报道。

鉴于此,本文借鉴现有研究成果,提出一种基于深度卷积神经网络的水稻知识文本分类方法。基于 ResNet[24]和 Inception V[25]网络结构的基本原理,以 Top－1 准确率为判断标准,分别从网络模块结构和网络层次进行分析,筛选具有最佳特征提取性能的 CNN 网络结构,以提高精准率、召回率、F1 值和正确率为目标,将筛选出的 CNN 网络结构与 CapsNet 相结合,建立水稻知识文本分类模型,以期为水稻知识文本的精准分类提供科学和理论依据。

1 语料采集及预处理

1.1 语料采集

本文通过采用 Python 爬虫框架,爬取知网专家在线系统和种植问答网等关于水稻病虫害、草害药害以及栽培管理等中文文本问答数据。同时,对所获数据进行初步人工筛选,最终获得 14 527 条有效数据,其中水稻病虫害、草害药害、栽培管理和其他数据分别为 5 640、1 335、6 060、1 492 条。水稻知识数据主要用于文本分类网络的模型训练与测试,每次试验从数据集中随机抽取 80% 作为训练集,10%

作为验证集，剩余 10% 作为测试集。

1.2　语料预处理

1.2.1　文本分词处理

相对英文文本，中文文本的处理相对复杂。中文字与字之间没有间隔，并且单个汉字具有的意义也明显弱于词组，因此本研究采用 Jieba 方法[26]对水稻知识文本进行分词处理，并去除文本中无用符号和停用词等。与此同时，中文分词结果深受分词词库的影响，为提高水稻知识文本分词精度，减少错分、漏分和误分情况，本文在搜狗农业语料库基础上构建水稻相关语料库，进而扩大 Jieba 分词基础词库，提高对水稻病虫害、草害药害和栽培管理等专业词汇的辨识度。

1.2.2　文本向量化处理

由于网络模型无法对自然语言进行直接训练学习，并且中文文本语句中存在大量的语义信息、上下文依赖信息和语序信息等，直接采用中文文本将无法保留这些信息的完整性，因此将中文文本转换为多维且连续的向量至关重要。本研究采用 Word2Vec[27]的 Skip－Gram 模型对水稻知识文本进行向量化处理。

2　水稻知识文本分类网络设计

2.1　ResNet－18 网络结构

建立特征提取层网络是解决文本分类问题的基础，而 CNN 在图像和文本特征提取问题上取得了较好的提取精度。ResNet 是 CNN 的典型代表，其残差模块（包括直接映射和残差部分）的设计理念使得随着网络层数的增加，网络发生退化的现象得以解决，且在 ILSVRC 2015 竞赛中其分类和特征提取的效果上得到了充分肯定[28]。图 1 为适用于文本分类的 ResNet－18 网络结构。其中 n Conv1D，m 等表示尺寸为 n、通道数为 m 的卷积核；Maxpooling1D，/2 表示步长为 2 的最大池化层；FC，4 表示通道数为 4 的全连接层。

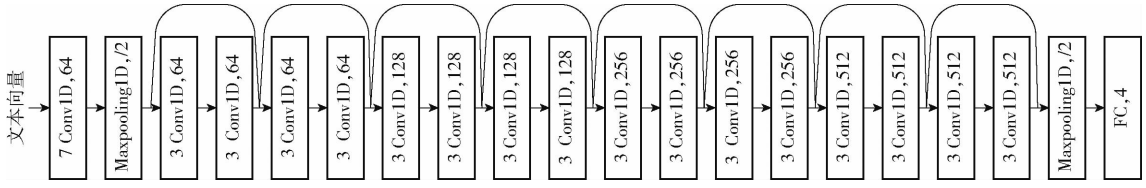

图 1　ResNet－18 网络结构

Fig. 1　ResNet－18 network architecture

ResNet－18 网络多用于图像分类，其采用多个 3×3 的二维卷积核 Conv2D 从图像矩阵的行维度和列维度进行特征提取，但文本向量是由规定长度的词向量按一定顺序构建的向量矩阵，所以从矩阵的行维度卷积（即从左至右移动）没有实际意义。因此采用多个尺寸为 n 的一维卷积核（Conv1D）仅从向量矩阵的列维度进行卷积。但由图 1 可知，ResNet－18 被采用一维最大池化方法（Maxpooling1D）的池化层分割为 3 部分，前两部分由 17 个尺寸为 7 和 3，通道数为 8、16、32 和 64 的卷积层构成，后一部分仅采用 1 个通道数为 4 的全连接层。显然直接将上述 ResNet 网络结构用于水稻知识文本特征提取并不合适。首先相比图像具有颜色和形状等规律性特征，水稻知识每一类数据均由几十个甚至更多的关键词组成，生成的文本向量具有一定复杂性，因此仅采用[3 Conv1D，3 Conv1D]结构的残差模块无法较为精准地提取文本特征。其次水稻知识文本向量具有较大离散性和稀疏性，直接采用 18 个权重层的 ResNet 网络结构易造成过拟合。鉴于此，本研究对 ResNet 的残差模块结构和网络层次进行设计与分析。

2.2　水稻知识文本分类设计

按照 Inception V 系列网络结构原理，面向卷积通道对 ResNet 的残差模块进行更改与设计。首先将 ResNet 的单通道卷积组调整为多通道卷积组，用以减少文本特征的表征性瓶颈（即减少信息损失），其次通过增加尺寸为 1 的卷积核对文本向量进行降维，并加入非线性，进而降低网络模型参数和提高网络的表达能力，因此共设计了 4 种结构的残差模块，如表 1 所示。

表 1　面向通道的 4 种残差模块结构

Tab. 1　Channel-oriented four residual module structures

种类	残差模块结构
A	$[1,\mathrm{Conv1D}]\times1,\begin{bmatrix}1,\mathrm{Conv1D}\\7,\mathrm{Conv1D}\\7,\mathrm{Conv1D}\end{bmatrix}\times1$
B	$\begin{bmatrix}3,\mathrm{Conv1D}\\1,\mathrm{Conv1D}\end{bmatrix}\times1,\begin{bmatrix}1,\mathrm{Conv1D}\\7,\mathrm{Conv1D}\\7,\mathrm{Conv1D}\\3,\mathrm{Conv1D}\end{bmatrix}\times1$
C	$[1,\mathrm{Conv1D}]\times1,\begin{bmatrix}1,\mathrm{Conv1D}\\3,\mathrm{Conv1D}\end{bmatrix}\times1,\begin{bmatrix}1,\mathrm{Conv1D}\\3,\mathrm{Conv1D}\\3,\mathrm{Conv1D}\end{bmatrix}\times1$
D	$[1,\mathrm{Conv1D}]\times1,[3,\mathrm{Conv1D}]\times1,[5,\mathrm{Conv1D}]\times1,[5,\mathrm{Conv1D}]\times1$

注：[3，Conv1D]表示尺寸为 3 的卷积核，下同。

为对比残差模块结构对文本分类的影响，共配置了 4 种水稻知识文本分类网络，如表 2 所示，并通过后续试验分析，筛选分类性能较高的残差结构。在保持较优残差结构不变的前提下，通过增加残差模块数量探究网络层次对分类精度的影响。

表 2 基于 4 种残差模块的网络结构

Tab. 2 Network structure based on four kinds of residual modules

网络类型	网络结构
A－NN	Embedding－A－Maxpool/2－FC/128－FC/4－Softmax
B－NN	Embedding－B－Maxpool/2－FC/128－FC/4－Softmax
C－NN	Embedding－C－Maxpool/2－FC/128－FC/4－Softmax
D－NN	Embedding－D－Maxpool/2－FC/128－FC/4－Softmax

注：Embedding 表示嵌入层；A 等表示残差模块类型，下同；Maxpool/2 表示步长为 2 的池化层；FC/128 表示通道数为 128 的全连接层，下同。

3 试验结果与分析

3.1 文本向量化处理与分析

采用 Word2Vec 中的 Skip－Gram 模型对水稻知识文本进行向量化处理，词向量维度为 100，训练窗口尺寸设置为 5。同时与 One－Hot[29]、TF－IDF[30]、Hashing[31]向量化模型进行对比分析。对 4 种模型训练得到的文本向量进行浅层神经网络建模，其精准率、召回率和 F1 值的宏平均值和正确率如表 3 所示。

表 3 4 种文本向量化建模结果

Tab. 3 Results of four kinds of text vectorization modeling %

向量化方法	精准率宏平均值	召回率宏平均值	F1 值宏平均值	正确率
One－Hot	83.62	60.19	62.53	79.77
TF－IDF	62.43	46.02	42.42	72.54
Hashing	35.38	42.64	38.18	68.48
Word2Vec	86.81	72.60	76.99	86.44

由表 3 可知，4 种基于文本向量化方法构建的浅层神经网络中，Word2Vec 方法相比其他方法具有最高的分类精度，正确率为 86.44%，Hashing 方法的分类效果最差。这可能是由于 One－Hot 所产生的向量维度较高，存在稀疏性，影响了神经网络的分类效果，TF－IDF 和 Hashing 虽然考虑字词间的语义信息，但问题也较为明显，这 2 种方法没有解决向量维度高和数据稀疏的问题，并且随着提取连续字的集合的增大，维度将会变得更高。从每一类的分类效果来看（图 2），基于 4 种向量化方法的浅层神经网络在栽培管理和病虫害上分类效果较好，在其他

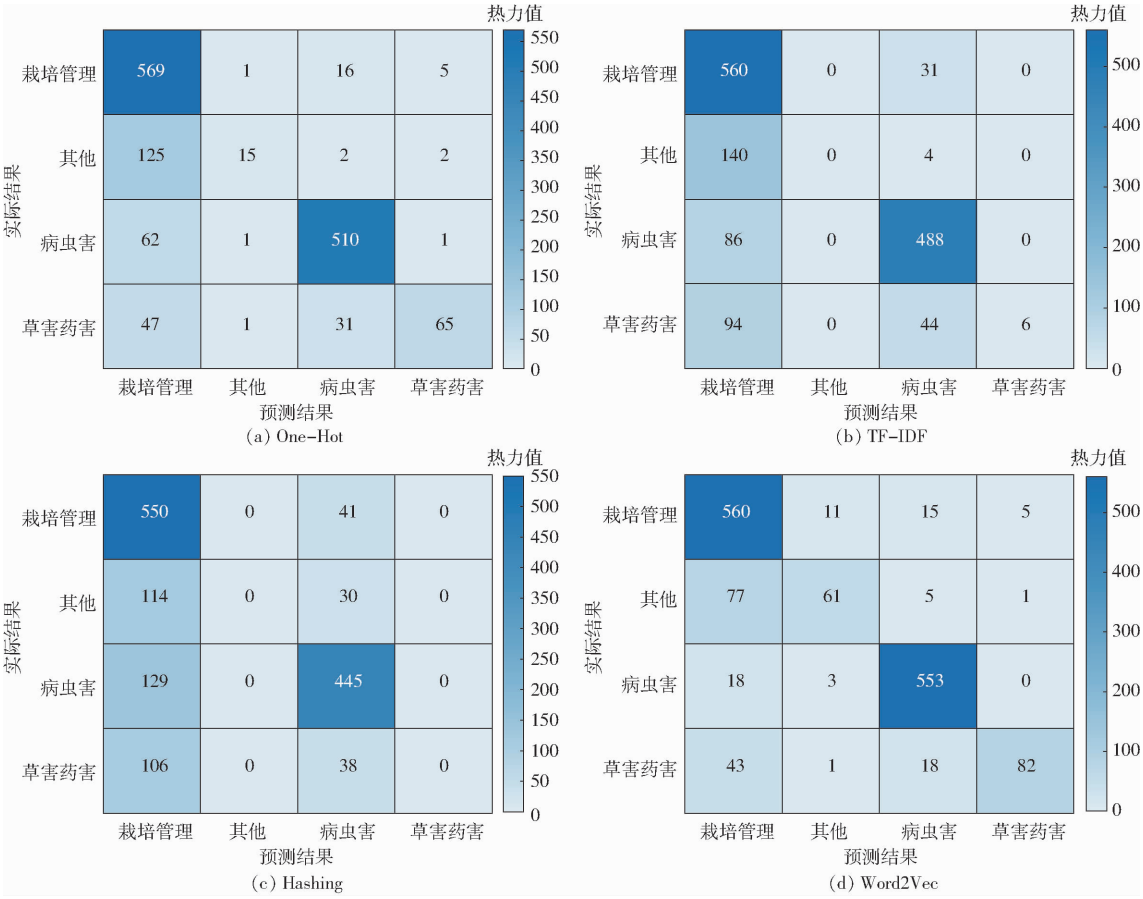

图 2 4 种文本向量化建模热力图

Fig. 2 Four kinds of text vectorization modeling heat maps

2 个类别上效果较差，原因在于草害药害和其他类别的数据量较小。但相较而言，Word2Vec 在草害药害和其他 2 个类别上向量化效果较好，网络模型能较为准确地提取到一定的文本特征，因此本研究采用 Word2Vec 模型构建文本向量作为后续研究的数据基础。

3.2　水稻知识文本分类模型训练与分析

采用 Word2Vec 生成的 4 类 14 527 条水稻知识文本向量作为样本，随机选取 80% 数据作为训练集，10% 作为验证集，并根据表 2 中 4 种模块结构，分别构建分类模型，开展相关试验与分析。采用 Top－1 准确率作为评价指标（Top－1 准确率为判断概率最大的类别与实际类别相符的准确率）。A－NN、B－NN、C－NN 和 D－NN 的 Top－1 准确率分别为 99.52%、99.11%、99.59% 和 99.38%。

可知，基于 4 种残差模块构建的网络模型均具有较好的分类精度，Top－1 准确率均达 95% 以上，其中残差模块 C-NN 所构建的分类模型具有最高的分类精度，Top－1 准确率为 99.59%，残差模块 D-NN、A-NN 和 B-NN 的分类性能逐渐降低。这可能是由于残差模块 C-NN 在各个通道卷积的第 1 层均采用了尺寸为 1 的卷积核，其能够在一定程度上增加非线性激励，提高了网络的表达能力，同时卷积通道数的增加使卷积核的数量增大，能够更充分地从数据中获取更多的文本特征。因此残差模块 C-NN 具有最佳的文本特征提取能力。与此同时，本研究在保持残差模块 C-NN 的基本结构不变的基础上，通过增加残差模块的数量（即增大网络深度）进行进一步训练与分析，结果如表 4 所示。

表 4　基于不同残差模块数量的网络分类性能

Tab.4　Network classification performance based on number of different residual modules

网络类型	网络结构	Top－1 准确率/%	残差模块数量
Ⅰ	Embedding－C×1－Maxpool/2－FC/128－FC/4－Softmax	99.59	1
Ⅱ	Embedding－C×2－Maxpool/2－FC/128－FC/4－Softmax	99.72	2
Ⅲ	Embedding－C×4－Maxpool/2－FC/128－FC/4－Softmax	99.79	4
Ⅳ	Embedding－C×6－Maxpool/2－FC/128－FC/4－Softmax	99.66	6
Ⅴ	Embedding－C×8－Maxpool/2－FC/128－FC/4－Softmax	99.24	8

注：C×2 表示连续 2 个残差模块 C，下同。

由表 4 可以看出，针对水稻知识文本样本，在保持残差结构相同的情况下，网络Ⅲ的分类效果最佳，Top－1 准确率为 99.79%，网络Ⅰ和Ⅱ的分类效果略差，说明当残差模块较少时，可适当增加模块数量，提高文本分类精度。但在网络Ⅲ的基础上，再增加残差模块数量时，网络的整体性能开始趋于饱和，分类精度有所下降。可能原因在于水稻知识数据中存在“共享词汇”，随着残差模块的增加，卷积数增大，模型训练得到一定共享词汇等的非主要文本特征，从而影响模型测试精度。

但是如果采用上述分类网络直接用于水稻知识文本分类，均需要采用池化层进行下采样操作。虽然池化层具有降低特征维度、减小模型参数等作用，但是文本特征经池化层操作后所得到的特征为标量，这将会导致文本特征向量的矢量方向信息和文本整体与词组之间的关联信息丢失（即文本的位置特征丢失），影响模型分类精度。而 CapsNet 采取向量进、向量出的训练模式，能够充分地保留文本向量特征。这与池化层的下采样操作截然相反。与此同时，CapsNet 首先采用 Squash 激活函数对特征向量进行压缩处理，保留向量的模长信息，从而能够表达特征向量所包含的信息强度。其次采用 Dynamic Routing（动态路由方法）对向量进行聚类分析，强化特征向量中的相似特征，弱化离群特征，输出更具有表达能力的文本特征。因此本研究采用胶囊网络（CapsNet）替代池化层，并结合前文的 4 层残差网络结构，构建基于深度卷积神经网络的水稻知识文本分类模型，简称为 RIC－Net，其网络架构如图 3 所示。

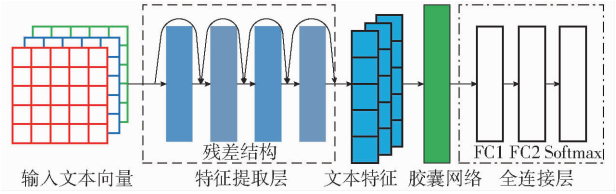

图 3　水稻知识文本分类模型网络结构

Fig.3　Network structure of rice knowledge text classification model

由图 3 可知，该网络将 Word2Vec 生成的文本向量作为模型输入，通过特征提取层提取文本特征，生成文本特征矩阵，并采用 CapsNet 替代池化层，对文本特征矩阵进行进一步筛选，最后经由全连接层和 Softmax 构成分类器，实现水稻知识文本的精准分类。

3.3　RIC－Net 训练与分析

在 RIC－Net 模型中，特征提取层的卷积滤波器数量分别为 8、16、32 和 64，CapsNet 的输出数量和

维度均为 50,动态路由轮数为 3,在全连接层中 FC1 和 FC2 神经元个数分别设置为 128 和 4。另外,采用 Nadam 算法(Nesterov-accelerated adam)[32]对模型进行训练,初始学习率为 0.002,一阶和二阶指数衰减率分别为 0.9 和 0.999。同时,经多次试验得出,经过 50 代训练,网络模型的训练损失均收敛到稳定值。为对比本文方法的分类效果,利用同一样本数据,分别训练 FastText、BiLSTM、Atten - BiGRU、RCNN、DPCNN 和 TextCNN 等 6 种常用文本分类模型,训练误差结果如图 4 所示。

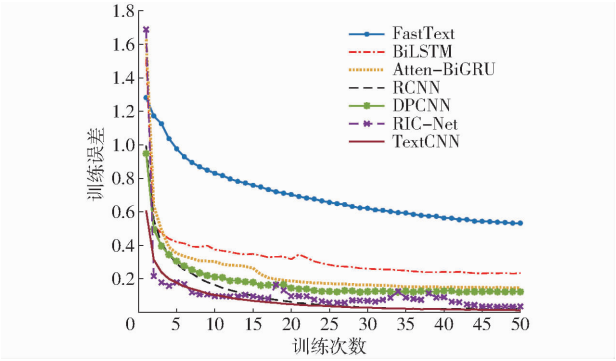

图 4 多种分类模型训练误差比较

Fig. 4 Comparison of training errors of various classification models

由图 4 可以看出,随着训练次数的增加,各个模型的训练误差均呈现不同程度的降低,当降低到一定程度后训练损失均收敛到稳定。在训练初始阶段,RIC - Net 的训练误差下降最快,说明 RIC - Net 能够较为精准地提取文本特征,模型更易收敛。同时,RIC - Net、TextCNN 和 RCNN 训练效果较好,当训练到 45 次时不仅达到收敛状态,且训练误差在 0 ~ 0.036 之间,基本达到训练要求。

3.4 RIC - Net 测试与分析

采用测试集分别对 RIC - Net 和其他 6 种文本分类方法进行测试与分析,并以精准率、召回率、F1 值以及正确率作为模型分类性能评价指标。评价结果如表 5 所示。

表 5 不同分类网络的测试结果比较

Tab. 5 Comparison of test results with different classification networks %

网络类型	精准率				召回率				F1 值				正确率
	类 1	类 2	类 3	类 4	类 1	类 2	类 3	类 4	类 1	类 2	类 3	类 4	
FastText	80.18	93.76	94.25	79.73	95.29	96.86	56.94	40.97	87.08	95.29	71.00	54.13	86.79
BiLSTM	95.59	97.92	93.71	85.42	95.09	98.60	93.06	85.42	95.34	98.26	93.38	85.42	95.32
DPCNN	98.42	98.10	95.21	86.62	95.26	98.78	96.52	94.44	96.82	98.44	95.86	90.37	96.70
RCNN	97.13	98.61	95.77	90.91	97.29	98.95	94.44	90.28	97.21	98.78	95.10	90.59	96.97
TextCNN	96.96	98.78	95.80	89.73	97.12	98.43	95.14	90.97	97.04	98.60	95.47	90.34	96.83
Atten - BiGRU	98.61	98.10	92.57	90.67	95.94	99.13	95.14	94.44	97.26	98.61	93.36	92.52	96.97
RIC - Net	98.98	99.13	95.58	95.17	98.98	99.48	96.53	95.83	98.98	99.30	97.54	95.50	98.62

注:类 1、类 2、类 3 和类 4 分别表示栽培管理、病虫害、草害药害和其他 4 个水稻知识类别。

由表 5 可知,与 FastText、BiLSTM、Atten - BiGRU、RCNN、DPCNN 和 TextCNN 等 6 种分类模型相比,RIC - Net 在栽培管理、病虫害、草害药害和其他类别上均具有较高的分类性能,对水稻知识的 4 种文本类型分类的精准率、召回率和 F1 值分别不小于 95.17%、95.83% 和 95.50%,在测试集的正确率方面,RIC - Net 同样高于其他模型,正确率为 98.62%。这是由于借鉴了 ResNet 和 Inception V 的基本思想,采用多通道和残差模块的结构构建特征提取层,精准提取水稻知识文本特征,同时将 CapsNet 替换池化层,保留了词组间相对位置特征,从而提高了模型分类精度。

从整体测试结果来看,相比数据量较大的病虫害分类结果,其他和草害药害的分类精度较低,说明增大类别样本量,能够进一步提高模型分类精度。但从数据量较小的其他和草害药害分类结果来看(图 5),RIC - Net 分类的精准率、召回率和 F1 值略高于其他 6 种模型。说明 RIC - Net 在样本量较少的情况下,也具有较高分类精度。

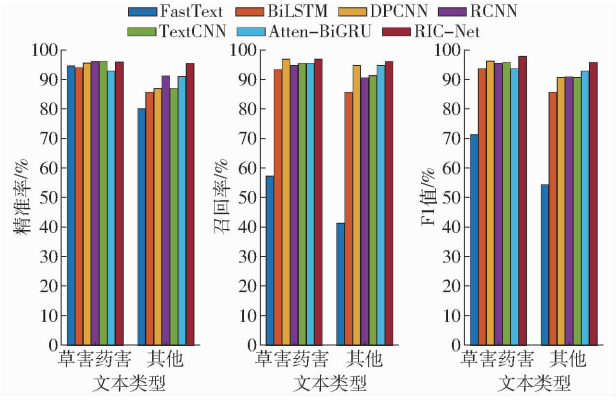

图 5 不同分类模型在草害药害和其他类别的对比

Fig. 5 Comparative analysis of different classification models in weeds damage categories and other categories

测试结果同时表明,相比栽培管理、病虫害和草

害药害的分类结果，7 种网络模型在其他类别上分类性能最低，原因在于在其他类别中存在多种繁杂的水稻相关知识点，因此该类文本中主要的关键词种类较多，缺乏一定的统一性，从而导致模型提取的特征不显著，降低了分类精度。但从图 6 可直观地看出，相比其他模型，RIC－Net 在其他类别中能够取得较好的分类结果，说明该模型具有较好的鲁棒性。

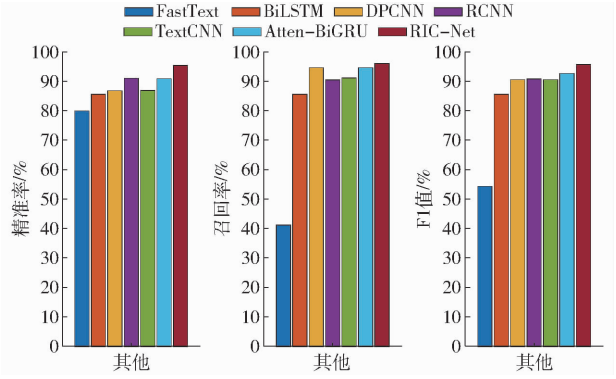

图 6　不同分类模型在其他类别的对比

Fig. 6　Comparative analysis of different classification models in other categories

4　结论

（1）相比 One－Hot、TF－IDF 和 Hashing 方法，采用 Word2Vec 方法对水稻知识文本数据进行向量化处理能够较好地保留文本语句中的语义和上下文依赖关系等信息的完整性。

（2）在 ResNet 的基础上，分别对其残差模块结构和网络层次进行改进设计，以便在对水稻知识文本分类过程中提高模型获取文本特征的能力。对 9 种分类网络结构的对比分析表明，采用 4 层 C 结构的残差模块作为水稻知识文本分类模型的基础网络结构能够较为精准地提取文本特征，Top－1 准确率达到 99.79%。

（3）本文所设计的水稻知识文本分类模型能够较好地对不同样本量和不同复杂程度的水稻知识数据进行精准分类。与 FastText、BiLSTM、Atten－BiGRU、RCNN、DPCNN 和 TextCNN 等 6 种模型相比，本文模型对水稻病虫害、草害药害、栽培管理和其他 4 种文本类别上均具有较高的分类性能，分类的精准率、召回率和 F1 值分别不小于 95.17%、95.83% 和 95.50%，正确率为 98.62%，满足实际应用需求。

参　考　文　献

[1] 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.

[2] DUAN Liguo, DI Peng, LI Aiping. A new naive Bayes text classification algorithm[C]// IEEE International Conference on Data Mining. IEEE, 2014: 947－952.

[3] JI L, CHENG X, KANG L, et al. A SVM-based text classification system for knowledge organization method of crop cultivation [C]// International Conference on Computer & Computing Technologies in Agriculture. Springer, Berlin, Heidelberg, 2011: 318－324.

[4] 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.
ZHOU Yuncheng, XU Tongyu, DENG Hanbing. Agricultural text classification method based on NB and CHI values[J]. Jiangsu Agricultural Sciences, 2018,46(17):219－223. (in Chinese)

[5] FENLIN W, YIFEI Z, CHENG W. Adaptive normalized weighted KNN text classification based on PSO[J]. Scientific Bulletin of National Mining University, 2016(1):109－115.

[6] 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 test 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)

[7] 段青玲，魏芳芳，张磊，等. 基于 Web 数据的农业网络信息自动采集与分类系统[J]. 农业工程学报，2016，32(12)：172－178.
DUAN Qingling, WEI Fangfang, ZHANG Lei, et al. Automatic acquisition and classification system for agricultural network information based on Web data[J]. Transactions of the CSAE, 2016, 32(12): 172－178. (in Chinese)

[8] 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? flag=1&file_no=20181226&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2018.12.026. (in Chinese)

[9] HUANG Ting, SHEN Gehui, DENG Zhihong. Leap-LSTM: enhancing long short-term memory for text categorization[EB/OL]. [2019－05－28]. https://arxiv.org/abs/1905.11558v1.

[10] REN Hao, LU Hong. Compositional coding capsule network with k-means routing for text classification[EB/OL]. [2018－10－22]. https://arxiv.org/abs/1810.09177

[11] ER M J, ZHANG Yong, WANG Ning, et al. Attention pooling-based convolutional neural network for sentence modelling[J].

Information Sciences,2016,373:388 – 403.
[12] WANG Haitao, HE Jie, ZHANG Xiaohong, et al. A short text classification method based on N – Gram and CNN[J]. Chinese Journal of Electronics,2020,29(2):248 – 254.
[13] 吴玉佳,李晶,宋成芳,等. 基于高效用神经网络的文本分类方法[J]. 电子学报,2020,48(2):279 – 284.
WU Yujia, LI Jing, SONG Chengfang, et al. High utility neural networks for text classification[J]. Acta Electronica Sinica, 2020,48(2):279 – 284. (in Chinese)
[14] 王文广,陈运文,蔡华,等. 基于混合深度神经网络模型的司法文书智能化处理[J]. 清华大学学报(自然科学版),2019,59(7):505 – 511.
WANG Wenguang, CHEN Yunwen, CAI Hua, et al. Judicial document intellectual processing using hybrid deep neural networks[J]. Journal of Tsinghua University(Science and Technology),2019,59(7):505 – 511. (in Chinese)
[15] 刘梓权,王慧芳,曹靖,等. 基于卷积神经网络的电力设备缺陷文本分类模型研究[J]. 电网技术,2018,42(2):644 – 651.
LIU Ziquan, WANG Huifang, CAO Jing, et al. A classification model of power equipment defect texts based on convolutional neural network[J]. Power System Technology,2018,42(2):644 – 651. (in Chinese)
[16] CHENG Hao, YANG Xiaoqing, LI Zang, et al. Interpretable text classification using CNN and max-pooling[EB/OL]. [2019 – 10 – 14]. https://arxiv.org/abs/1910.11236.
[17] 范丽丽,赵宏伟,赵浩宇,等. 基于深度卷积神经网络的目标检测研究综述[J]. 光学精密工程,2020,28(5):1152 – 1164.
FAN Lili, ZHAO Hongwei, ZHAO Haoyu, et al. Survey of target detection based on deep convolutional neural networks[J]. Optics and Precision Engineering,2020,28(5):1152 – 1164. (in Chinese)
[18] 周云成,许童羽,郑伟,等. 基于深度卷积神经网络的番茄主要器官分类识别方法[J]. 农业工程学报,2017,33(15):219 – 226.
ZHOU Yuncheng, XU Tongyu, ZHENG Wei, et al. Classification and recognition approaches of tomato main organs based on DCNN[J]. Transactions of the CSAE,2017,33(15):219 – 226. (in Chinese)
[19] 金宁,赵春江,吴华瑞,等. 基于 BiGRU_MulCNN 的农业问答问句分类技术研究[J/OL]. 农业机械学报,2020,51(5):199 – 206.
JIN Ning, ZHAO Chunjiang, WU Huarui, et al. Classification technology of agricultural question based on BiGRU_MulCNN [J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(5):199 – 206. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 20200522&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2020.05.022. (in Chinese)
[20] 贾旭东,王莉. 基于多头注意力胶囊网络的文本分类模型[J]. 清华大学学报(自然科学版),2020,60(5):415 – 421.
JIA Xudong, WANG Li. Text classification model based on multi-head attention capsule neworks[J]. Journal of Tsinghua University(Science and Technology),2020,60(5):415 – 421. (in Chinese)
[21] 唐贤伦,林文星,杜一铭,等. 基于串并行卷积门阀循环神经网络的短文本特征提取与分类[J]. 工程科学与技术,2019,51(4):125 – 132.
TANG Xianlun, LIN Wenxing, DU Yiming, et al. Short text feature extraction and classification based on serial-parallel convolutional gated recurrent neural network[J]. Advanced Engineering Sciences,2019,51(4):125 – 132. (in Chinese)
[22] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882,2014.
[23] YANG M, ZHAO W, YE J B, et al. Investigating capsule networks with dynamic routing for text classification [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL,2018: 3110 – 3119.
[24] ALEXIS C, HOLGER S, LOÏC B, et al. Very deep convolutional networks for text classification[EB/OL]. [2017 – 01 – 27]. https://arxiv.org/abs/1606.01781.
[25] JIANG W, JIN Z. Integrating bidirectional LSTM with inception for text classification[C]// IAPR Asian Conference on Pattern Recognition. IEEE Computer Society, 2017.
[26] CHEN Jiahao, ZHANG Jiayi. An industry classification model of small and medium-sized enterprises based on TF – IDF characteristics[C]// Proceedings of 2019 International Conference on Arts, Management, Education and Innovation (ICAMEI 2019),2019:273 – 277.
[27] ZHANG D, XU H, SU Z, et al. Chinese comments sentiment classification based on Word2Vec and SVMperf[J]. Expert Systems with Application, 2015, 42(4):1857 – 1863.
[28] HE Kaiming, ZHANG Xianyu, REN Shaoqing, et al. Deep residual learning for image recognition[EB/OL]. [2015 – 12 – 10]. https://arxiv.org/abs/1512.03385.
[29] JOHSON Rie, ZHANG Tong. Supervised and semi-supervised text categorization using One – Hot LSTM for region embeddings [EB/OL]. [2016 – 02 – 07]. https://arxiv.org/abs/1602.02373v1.
[30] TRSTENJAK B, MIKAC S, DONKO D. KNN with TF – IDF based framework for text categorization[J]. Procedia Engineering, 2014,69(1): 1356 – 1364.
[31] CHI Lianhua, LI Bin, ZHU Xingquan. Context-preserving hashing for fast text classification[C]// International Conference on Computer Networks & Mobile Computing. IEEE Xplore, 2014.
[32] LE T T H, KIM J, KIM H. An effective intrusion detection classifier using long short-term memory with gradient descent optimization[C]// International Conference on Platform Technology & Service. IEEE, 2017:1 – 6.