http://myfj.cnjournals.com myfj_sadri@163.com (021)20866796

DOI: 10.19416/j.cnki.1674-9804.2023.04.001

基于 TextCNN 的试飞运营问题分类

俞立群*

(上海飞机设计研究院,上海 201210)

摘 要: 针对需要识别海量试飞运营问题中的故障问题用于可靠性指标计算评估,基于深度学习中的文本卷积神经网络,提出一种试飞运营问题文本分类方法。通过收集大量的以人工分类的试飞运营问题文本作为实验数据集,并进行相应的预处理,运用 Word2Vec 模型将问题描述文本训练成词向量,构建出 TextCNN 模型进行训练完成问题文本的分类。最后通过实验表明,基于 TextCNN 模型的试飞运营问题分类方法可以为试飞运营问题自动化分类工作提供参考。

关键词: TextCNN;试飞运营问题;文本分类

中图分类号: TP391.1 **文献标志码**: A **OSID**:

0 引言

民用飞机可靠性指标是民用飞机最重要的评价指标之一。很多可靠性指标评估过程依赖于飞机故障数据的获取,如平均失效间隔时间(mean time between failures,简称 MTBF)和故障千时率等的计算都涉及故障次数。但是原始试飞运营问题清单特别是试飞问题清单中没有“类别”字段,也缺乏能直接支持分类的信息,导致无法直接开展可靠性指标评估工作。此外原始试飞运营问题清单中不仅包含了故障类数据,还包含了优化类和咨询类等非故障类数据,例如“APPR 模式,自动飞行的控制逻辑是什么?”是一条咨询类数据。因此,需要根据问题数据的文本描述进行分类。此外,原始试飞运营问题清单数据众多而且非故障类数据也占有不小比例,那么进行问题分类后能让工程师更加快速聚焦于故障问题。

文本分类是自然语言处理(nature language process,简称 NLP)领域中一项重要任务,已经有许多学者提出不同的方法完成文本分类任务,主要有两类。一类是传统机器学习方法,包括 SVM[1]、朴素贝叶斯[2]、KNN[3]和基于规则特征的方法[4]。另一类是基于深度学习的多种神经网络被应用于文本分类任务中,考虑序列的循环神经网络(recurrent neural network,简称 RNN)模型[5]。把注意力模型引入序列模型中也对分类效果带来了较大提升[6]。卷积神经网络(convolutional neural networks,简称 CNN)[7]被广泛用于处理计算机视觉任务当中,随着词嵌入和深度学习技术的发展,Collobert[8]首次将 CNN 模型应用到 NLP 领域中,Kim[9]则第一个将 CNN 模型应用于文本分类的任务当中,提出了经典的 TextCNN 模型。ZHANG Ye[10]对 TextCNN 模型的参数设置的敏感性进行了分析。Johnson[11]则修改了 TextCNN 的文本嵌入方法,提升了文本情感分类任务的效果。KALCHBRENNER[12]将池化层的值设置为动态,即在不同层级的网络中设置不同的值。韩栋[13]等人在句子级层面融合主题句与 CNN,提出句子级 CNN 监督学习文本分类方法。

在 CNN 的基础上,TextCNN 模型被提出用于文本分类工作。TextCNN 的优点在于根据卷积核大小的不同实现对局部特征的提取,从而使提取到的特征向量具有多样性且更具代表性。经典 TextCNN 模型结构清晰、计算效率高,因此被广泛使用于工程实际中,在文本二分类任务上也取得了很好的结

* **通信作者**. E-mail: yuliqun@comac.cc

引用格式: 俞立群. 基于 TextCNN 的试飞运营问题分类[J]. 民用飞机设计与研究,2023(4):1-5. YU L Q. Classification of flight test and operation problems based on TextCNN[J]. Civil Aircraft Design and Research, 2023(4): 1-5 (in Chinese).

果。考虑到民用飞机设计领域存在许多专有词汇,因此,在经典 TextCNN 模型的基础上,引入民用飞机设计领域专有词汇库,对试飞运营问题进行分类。

1　TextCNN 模型

TextCNN 模型利用多个不同大小的卷积核实现对文本信息的关键特征进行提取。TextCNN 的模型结构如图 1 所示,可分为四层,首先是嵌入层,接着为卷积层和激活层,然后是池化层,最后是全连接层。

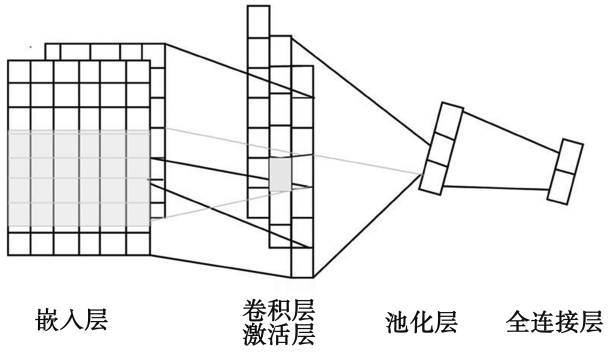

图 1　TextCNN 网络结构

1.1　嵌入层

嵌入层作为整个 TextCNN 网络的输入,输入形式为 n×k 的矩阵,每行为词向量,n 为词的数量,k 为规定的维度。词向量的训练首先经过分词,然后利用 Word2Vec 将词语转换成词向量,这个过程也称为词嵌入(word embedding)。例如首先将“右侧驾驶座椅电动调节失效”分词为“右侧/驾驶/座椅/电动/调节/失效”,再通过 word2Vec 方法映射成一个 5 维词向量(维度可取任意值),如式(1)和图 2 所示。构建完所有的词向量后,将它们拼接成一个 n×k 的矩阵作为 TextCNN 网络的输入。

右侧	0	0	0	0	1
驾驶	0	1	0	0	0
座椅	0	0	1	0	0
电动	1	0	0	0	0
调节	0	0	0	1	0
失效	1	0	1	0	0

图 2　词向量表示

$$W = \{x_1, x_2, x_3, \cdots, x_n\} \tag{1}$$

式中:W 表示由词向量组成的矩阵;x_i 表示第 i 个词的词向量。

1.2　卷积层和激活层

将输入的词向量矩阵与卷积核进行卷积运算操作,得到特征图(feature map),再对特征图进行一个非线性函数的映射。卷积核长度可以任意设置,宽为词向量的维度,可以设置多个卷积核,采用不同长,同一宽。卷积的核心公式如式(2)所示,在输入为 n×k 的矩阵上,使用一个卷积核与一个窗口进行卷积操作,卷积操作即词向量矩阵与卷积核分别对应相乘再相加。从上往下一次移动一步,每次进行卷积操作后得到一个特征向量 c_i,将这些特征向量拼接起来就得到特性图,如式(3)所示。卷积操作是一个线性计算,因此还需要利用激活函数对结果进行非线性变换,采用的激活函数为 Relu 函数。

$$c_i = f(w \times W_{i:i+h-1} + b) \tag{2}$$

$$c = [c_1, c_2, \cdots, c_{n-h+1}] \tag{3}$$

式中:c_i 表示第 i 个特性向量;f 为激活函数;W 表示权重矩阵;h 表示窗口中的单词数;$W_{i:i+h-1}$ 表述输入矩阵的第 i 行到 $i+h-1$ 行组成的 $h \times k$ 的窗口;b 表示偏置参数;c 表示特征图。

1.3　池化层

池化层紧跟在卷积层和激活层之后,作用在于通过删减数据,减少过拟合现象发生。采用 1-Max 最大池化方法(max-pooling),即取每一个特征向量中的最大值,最后把所有池化后的数据按行级联,得到最后的特征表达。

$$c_{\max} = \max(c_i) \tag{4}$$

式中:$c_{\max}$ 表示特征向量中最大值。

1.4　全连接层

将特征表达输入最后全连接层,为了防止过拟合,在倒数第二层的全连接部分上使用 dropout(暂退法)技术,最终可以将 N 维向量压缩到分类类别个数的维度,并输出每个类别的概率,选择最大的概率完成分类任务。

2　试飞运营问题分类

2.1　整体流程

试飞运营问题分类任务主要流程包括试飞运营问题数据收集、试飞运营问题文本数据预处理、试飞运营问题文本数据向量化表示、词向量矩阵输入 TextCNN 模型训练、选择概率最大的类别作为分类结果输出。

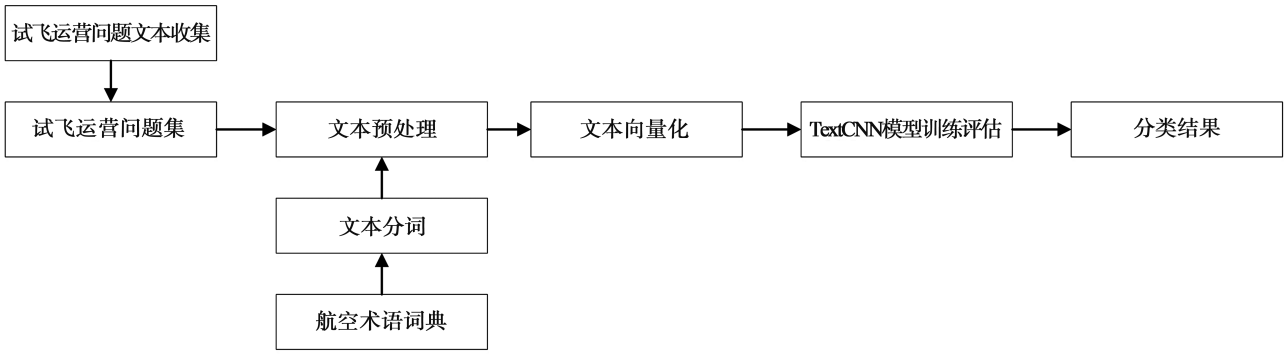

图 3　试飞运营问题分类流程

2.2　试飞运营问题收集与预处理

收集某型号在试飞运营过程中记录的问题描述文本数据,预处理步骤包括去重、清洗和重排序。由于 TextCNN 模型是有监督学习模型,所以训练数据需要分类标签。因此还需要对试飞运营问题打上“故障类”和“非故障类”标签用于后续模型的训练。

2.3 文本转换词向量

文本转换词向量的目的是将试飞运营问题文本转换成词向量。首先,利用 jieba 工具包对试飞运营问题文本进行分词,考虑到航空领域有众多专有词汇,在进行分词时加入航空术语词典以进一步提高分词准确率。然后,在得到分词结果后,利用停用词表去除停用词,一般后续还会去除词频较低的词,但考虑到航空领域一些词虽然低频但具有独特且重要的含义,能很好地表达句子的意思,因此本文省略去除低频词步骤。接着,利用已经准备好的词库文件,即词和数字的字典,将试飞运营问题文本分词后的结果转换为对应的数字 id。最后,若文本长度超过设定值,则将其截断保留后半部分;若长度不足,则前面补 0。最终得到试飞运营问题文本词向量表作为 TextCNN 模型的输入。

2.4　TextCNN 模型构建

首先,分别使用不同卷积核长度的词窗口去执行卷积,并采用 Relu 函数进行激活。然后,进行最大池化处理。接着,合并三个经过卷积和池化后的输出向量。最后,将合并后的输出向量输入全连接层,并在全连接层 1 和 2 之间添加 dropout 减少训练过程中的过拟合,随机丢弃一定比例的结点值。

2.5　评价指标

试飞运营问题分类结果评价指标为准确率 A、精确率 P、召回率 R 和 F_1。

准确率 A 的计算公式为:

$$A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{5}$$

式中:T_P 表示预测为正、实际为正的数量;T_N 表示预测为负、实际为负的数量;F_P 表示预测为正、实际为负的数量;F_N 表示预测为负、实际为正的数量。

精确率 P 的计算公式为:

$$P = \frac{T_P}{T_P + F_P} \tag{6}$$

召回率 R 的计算公式为:

$$R = \frac{T_P}{T_P + F_N} \tag{7}$$

F_1 值是精确率 P 和召回率 R 的调和平均数,其计算公式为:

$$F_1 = \frac{2PR}{P + R} \tag{8}$$

3　实例验证

3.1　数据集构建

飞机试飞运营问题数据集来源于某型号飞机试飞和运营过程中记录的问题文本描述。首先,由于数据来源渠道众多,存在重复问题,因此需要对获取的数据进行去重。其次,对数据进行清洗,包括去除前后空格和无意义不明字符等。由于原始数据记录具有很明显的时间序列特征,还需要将数据随机打乱,最终得到了 7 193 条数据。然后,对数据进行分类,分为“故障类”和“非故障类”,最终得到包含“标签”和“问题描述”两个字段的数据集。最后,按照 8∶2 的比例将数据切分成训练集和测试集。随着后续试飞和运营的开展,问题记录数量也会不断增加,可以通过对模型不断迭代训练,提高模型分类准确性。

3.2　参数设置

通过设置不同卷积核长度组合和 dropout 值,比较评价指标以探索参数设置对结果的影响程度和寻找相对较优的参数设置。

在卷积层中设置不同的卷积核长度并进行对比,选择效果最佳的卷积核组合。不同组合的评价指标对比情况如表 1 所示。

表 1　卷积核组合对比结果

卷积核组合	准确率	精准度	召回率	调和平均数
2、3、4	0.792 0	0.954 1	0.980 5	0.967 1
3、4、5	0.776 5	0.957 1	0.980 7	0.968 7
2、3、5	0.776 5	0.957 0	0.981 2	0.968 9
2、4、5	0.790 5	0.958 1	0.980 5	0.969 2

由实验结果可知,不同卷积核对实验结果影响灵敏度不高,相对来说设置卷积核大小分别为 2、4、5 时分类效果较好,故在试飞运营问题分类任务中设置卷积核大小分别为 2、4、5。

通过设置不同的 dropout 值,比对分类效果,评价指标结果如表 2 所示。

表 2　全连接层 dropout 值结果

dropout 值	准确率	精准度	召回率	调和平均数
0.00	0.780 5	0.956 0	0.982 2	0.968 9
0.15	0.790 3	0.957 1	0.982 0	0.969 4
0.25	0.790 5	0.958 1	0.980 5	0.969 2
0.35	0.781 5	0.953 3	0.980 6	0.966 7

由实验结果可知,不同的 dropout 值对实验结果影响灵敏度不高,在试飞运营问题分类任务中,dropout 值为 0.25 时分类效果较好。

4　结论

1）针对需要识别海量试飞运营问题中的故障问题,提出基于 TextCNN 模型的文本分类方法,用于根据试飞运营问题文本描述对故障类和非故障类问题进行分类,实验结果表明该方法可以为试飞运营问题自动化分类工作提供参考。

2）在文本预处理和向量化过程中,需加入航空术语词典以及省略去除低频词步骤。

3）不同卷积核长度和 dropout 值对实验结果灵敏度不高,相对来说试飞运营问题分类任务中卷积核长度设置为 2、4、5 以及 dropout 值取 0.25 时分类效果较好。

参考文献:

[1] 张华鑫. 基于 SVM 的文本分类研究[J]. 情报探索, 2015,34(5):133-135.

[2] WANG S D, MANNING C D. Baselines and bigrams: simple, good sentiment and topic classification[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, July 8-14, 2012, Jeju Island, Republic of Korea. [S. l.: s. n.], 2012:90-94.

[3] 卜凡军, 钱雪忠. 基于向量投影的 KNN 文本分类算法[J]. 计算机工程与设计, 2009, 30(21): 4939-4941.

[4] 靳义林, 胡峰. 基于三支决策的中文文本分类算法研究[J]. 南京大学学报(自然科学), 2018, 54(4): 794-803.

[5] 刘腾飞,于双元,张洪涛,等. 基于循环和卷积神经网络的文本分类研究[J]. 软件, 2018, 39(1): 64-69.

[6] 邓朝阳, 仲国强, 王栋. 基于注意力门控图神经网络的文本分类[J]. 计算机科学, 2022, 49(6): 326-334.

[7] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, 1(4): 541-551.

[8] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12:2493-2537.

[9] KIM Y. Convolutional neural networks for sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar. [S. l.: s. n.], 2014: 1746-1751[2023-08-28]. https://arxiv.org/pdf/1408.5882.pdf.

[10] ZHANG Y, WALLACE B C. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [C]//Proceedings of the 8th International Joint Conference on Natural Language Processing, November 27-December 01, 2017, Taipei, Taiwan. arXiv preprint arXiv: 1510.03820, 2015[2023-08-28]. https://aclanthology.org/I17-1026.pdf.

[11] JOHNSON R, ZHANG T. Semi-supervised convolutional neural networks for text categorization via region embedding[J]. Advances in neural information processing systems, 2015, 28: 919-927.

[12] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM

P. A convolutional neural network for modelling sentences[C/OL]//Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics, June 23-25, 2014, Baltimore, Maryland. arXiv preprint arXiv:1404.2188, 2014[2023-8-28]. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.636.1565&rep=rep1&type=pdf.

[13] 韩栋,王春华,肖敏. 基于句子级学习改进 CNN 的短文本分类方法[J]. 计算机工程与设计, 2019,40(1): 256-260,284.

作者简介

俞立群　男,硕士,工程师。主要研究方向:飞机数据管理与分析。E-mail: yuliqun@comac.cc

Classification of flight test and operation problems based on TextCNN

YU Liqun *

(Shanghai Aircraft Design and Research Institute, Shanghai 201210, China)

Abstract: Aiming to identify fault issues in a large amount of flight test and operation data for reliability indicator calculation and evaluation, a text classification method based on convolutional neural networks (CNN) in deep learning was proposed. By collecting a large amount of manually classified flight test and operation problem texts as the experimental datasets and performing corresponding preprocessing, the Word2Vec model was used to train the problem description text into word vectors, and a TextCNN model was constructed for training to complete the classification of problem texts. Finally, the experiments show that the classification method for flight test and operation problem based on TextCNN model can provide reference for the automated classification of flight test and operation problems.

Keywords: TextCNN; flight test and operation problems; text classification

* Corresponding author. E-mail: yuliqun@comac.cc