软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(5):1375−1385 [doi: 10.13328/j.cnki.jos.005730]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于维修日志的飞机设备故障原因判别方法*

王锐光, 吴 际, 刘 超, 杨海燕

(北京航空航天大学 计算机学院,北京 100191)

通讯作者: 吴际, E-mail: wuji@buaa.edu.cn

摘 要: 在飞机维修与保养过程中,航空维修公司已积累了大量经验性的维修日志数据.合理利用该类维修日志,结合机器学习方法,可以辅助维修人员做出正确的故障诊断决策.首先,针对维修日志的特殊性,提出一种迭代式的故障诊断基本过程;其次,在传统的文本特征提取技术的基础上,基于领域内信息,提出一种基于卷积神经网络(convolution neural network,简称 CNN)的小样本文本特征提取方法,在样本量较少的情况下,利用预测目标将字向量作为输入,得到更为充分的文本特征;最后,使用随机森林(random forest,简称 RF)模型,结合其他故障特征判别飞机设备的故障原因.卷积神经网络以故障原因为目标,预先对故障现象中的字向量进行训练,从而得到更能反映该领域的文本特征.与其他文本特征提取方法相比,该类方法在小样本数据上得到了更好的效果.同时,将卷积神经网络与随机森林模型应用于飞机设备的故障原因判别,并与其他文本特征提取方式和机器学习预测模型进行对比,说明了该类文本特征提取方式和故障原因判别方法的合理性和必要性.

关键词: 故障诊断;维修日志;卷积神经网络;随机森林

中图法分类号: TP311

中文引用格式: 王锐光,吴际,刘超,杨海燕.基于维修日志的飞机设备故障原因判别方法.软件学报,2019,30(5):1375−1385. http://www.jos.org.cn/1000-9825/5730.htm

英文引用格式: Wang RG, Wu J, Liu C, Yang HY. Fault cause identification method for aircraft equipment based on maintenance log. Ruan Jian Xue Bao/Journal of Software, 2019,30(5):1375−1385 (in Chinese). http://www.jos.org.cn/1000-9825/5730.htm

Fault Cause Identification Method for Aircraft Equipment Based on Maintenance Log

WANG Rui-Guang, WU Ji, LIU Chao, YANG Hai-Yan

(School of Computer Science and Engineering, BeiHang University, Beijing 100191, China)

Abstract: In the process of aircraft maintenance, the aviation maintenance company has accumulated a large number of empirical maintenance log data. Machine learning methods can be used to help maintenance staff to make correct fault diagnosis decisions, using this type of maintenance log reasonably. Firstly, according to the particularity of the maintenance log, an iterative fault diagnosis process is proposed. Secondly, based on the traditional text feature extraction technology, the text feature extraction method based on convolution neural network (CNN) with the information in the domain is proposed, which is used in the case of small sample size. The method uses the target vector to train word vector to get more adequate text features. Finally, the random forest (RF) model is used in combination with other fault characteristics to determine the cause of aircraft equipment failure. The convolutional neural network aims at the cause of the failure, and pre-trains the word vector in the fault phenomenon to obtain a text feature that better reflects the field. Compared with other text feature extraction methods, the method obtains better results in the case of small sample size. At the same time, the convolutional neural network and random forest model are applied to the identification of aircraft equipment failure, and compared with other text feature extraction methods and machine learning prediction models, which illustrates the rationality and necessity of the method of text feature extraction and the method of fault cause identification.

Key words: fault diagnosis; maintenance log; convolutional neural network; random forest

---

* 本文由智能化软件新技术专刊特约编辑申富饶教授和李戈副教授推荐.
收稿时间: 2018-09-01; 修改时间: 2018-10-31; 采用时间: 2018-12-13

众所周知,许多安全关键系统变得规模化、复杂化和高度耦合化,如航空发动机、汽车车辆、化学系统、电力系统、风能转换系统和工业电子设备等等.所以,对可能存在工艺异常和设备故障的系统,其可靠性和安全性的要求越来越高.简单的异常可能会损坏部分功能,从而造成经济损失甚至巨大的人员伤亡,故尽可能早地检测和识别潜在异常并实施容错操作以最小化性能降级和避免危险情况是至关重要的.随着航空业的飞速发展,航空公司的飞行安全需求不断提高,但飞机结构愈加复杂,同一故障可能由多种因素引起,设备之间关联的多变性,使得维修人员难以通过传统的基于故障诊断规则[1]和基于专家系统的故障诊断方式[2]得到准确的结论,更先进的监控手段和故障诊断技术应逐渐应用到复杂系统中.

1 引 言

故障诊断的目标是提供关于故障更加详尽的描述信息,包括但不限于故障检测、故障原因判断、故障定位及故障恢复等[3].一旦检测到故障,维修人员就需要根据经验判断故障原因,从而提出故障修复方案.故障原因是故障诊断的首要目标,确定故障原因之后才能进行相应的故障排除措施,及时地避免更大的经济损失和伤亡.

由于飞机等系统的结构、性质和先验知识难以在短期内获得,基于知识和基于模型的故障诊断难以进行下去.而使用基于数据驱动的故障诊断方法不需要完整的系统模型,只要求可靠的定量或定性数据,这使得该类故障诊断方法变得切实可行.许多学者已经提出几种性能较好的基于数据的故障诊断方法,如:文献[4]首先利用自助重采样方法对原始样本进行处理,基于不同样本的自助子集分别去训练不同的神经网络,最后对所有网络的诊断结果进行综合,从而提高了故障诊断的可靠性;文献[5]在配电系统的故障诊断中利用主成分分析法对训练样本进行降维,然后利用支持向量机和神经网络方法实现故障分类,达到了较好的分类精度;文献[6]研究了以复杂工业过程为重点的故障分类问题,为了进行多故障分类,研究了基本的支持向量机以及主成分分析方法,实验表明:标准主成分分析法仍然有令人满意的结果,而且计算量较少.

近年来,航空维修企业已经有相当规模的维修经验数据积累,大部分企业将该类维修经验用于专家系统的构建中[7].航空维修企业的数据包括非结构化数据和结构化数据两种,其中:结构化数据容易直接用于计算和分析;而非结构化数据可以用于分析,也可以通过自然语言处理等方法转换为结构化数据.Chiu C 等人[8]提出了基于案例推理的方法,使用历史非结构化维修案例数据,并采用遗传算法增强相似性函数性能的方法来检索相似案例,达到了较好的效果.李青等人[9]开发了基于案例推理和分词替换的故障诊断系统,通过标准词典的词条替换,将人为描述转换为更标准的格式,使语义类似的案例达到更高的相似度.文献[10]采用主题模型对高铁车载设备故障文本信息进行特征提取,基于贝叶斯网络对故障进行分类,达到了较好的诊断准确性.文献[11]中,针对汽车领域在故障诊断期间形成的大量文本数据,提出一种基于本体的文本挖掘技术的知识发现方法,使用诊断本体来发现最佳的实践经验以用来修复知识,该方法在现实工业中的基于 Web 的分布式架构中成功应用.文献[12]中,针对铁路维修部门的故障文本数据,提出了基于双层特征提取的文本挖掘方法,在语法层次上使用基于卡方统计的特征选择来解决样本不均衡问题,之后,在语义层次上使用基于 Dirichlet 分配的特征选择,以将数据降维至低维主题空间,并通过铁路公司收集的铁路维护数据验证了其性能.Zhao 等人[13]提出了一种基于文本挖掘技术的铁路车载设备故障诊断方法.该方法使用主题模型从维修记录中提取故障特征,同时采用贝叶斯网络调整故障诊断的不确定性和复杂性,最后,充分利用专家知识和数据以推导出合适的贝叶斯网络结构.该方法通过武广高速铁路信号系统的实际数据验证了正确性.

在文本特征提取方面,传统的方式一般有词袋模型(bag of words,简称 BOW)或向量空间模型(vector space model).除此之外,在特征权重方面,主要是经典的 TF-IDF[14]以及其他扩展方法.词袋模型的最大问题是维度和稀疏性很高,词与词之间相互独立,忽视了上下文关系,因此需要特征选择、降维等方法降低维度,通过特征权重增加稠密性.而向量空间模型虽然克服了词袋模型在高纬度上的缺点,但训练该类模型需要庞大的语料才能很好地反映词与词之间的上下文关系.本文首先提出一种迭代式的故障诊断基本过程,通过不断积累维修日志,提高故障诊断的准确度;其次,在传统文本特征提取技术的基础上,基于领域内信息,提出一种基于卷积神经网络的字符级文本特征提取方法,在样本量较少的情况下,取得了较好的效果;最后,使用随机森林模型结合其他故

障特征判别飞机设备故障原因,从而达到了较好的故障原因分类精度.

本文第 2 节介绍随机森林模型的理论基础.第 3 节介绍基于维修日志的故障诊断基本过程.第 4 节提出基于卷积神经网络的小样本字符级文本特征提取方法.第 5 节设计实验验证随机森林算法的优越性,说明本文方案的有效性.第 6 节对本文工作进行总结并提出后续研究方向.

2 随机森林

随机森林(random forest,简称 RF)[15]是基于多决策树的 Bagging 类集成学习算法,通过自助(bootstrap)重采样技术且并行训练多个基分类器来降低学习算法的方差,从而得到良好的分类性能.随机森林算法在故障分类领域中应用较多[16−18],原因主要有:参数数量较少,不需要大量的调参工作;由于 Bagging 的集成思想,所以不必担心过拟合现象的发生;对缺失值较多的数据能够很好地适用;能通过训练得到特征的重要程度;作为树结构,对多分类任务有良好的适应性;对于文本等高维数据具有良好的处理能力等.由于飞机维修日志经过结构化以后维度较高、且缺失值较多,所以采用随机森林算法作为主要的故障原因判别方法.

随机森林由所有决策树经过投票决定每个输入样本 X 的类别.每棵决策树$\{h(x,\theta_i),i=1,2,\ldots,k\}$依赖于$\theta_i$,且$\theta_i$是独立同分布的随机向量.而生成每棵决策树时的随机性,使得整体的泛化误差既依赖于单棵树的分类性能,也依赖于各决策树之间的相关关系.随机森林算法主要分为决策树的生成和随机森林投票两个步骤.

2.1 决策树生成

决策树分类是一种从杂乱无章的数据集中学习出树状表示形式的分类规则的方法.随机森林使用 CART 分类树作为基决策树,使用自助重采样技术生成每一棵决策树分类器.单棵决策树的生成过程如下描述[19].

1) 对原始训练集,使用有放回抽样的方式随机抽取训练样本,每个训练集大小约为原始训练集的 2/3.
2) 为每个 bootstrap 训练集建立 CART 决策树,一共产生 n_t 棵决策树,从而构成一片“森林”.
3) 随机选择数据集中的特征.假设训练数据集中有 M 个特征,从中随机选择 $m(m<M)$个特征.在 m 个特征中选择基尼指数最小的特征及其对应的切分点作为当前最优特征及最优切分点,并从该节点分裂为两个子节点;之后,对子节点递归进行上述步骤构造分支,直至该决策树能准确分类所有训练集或没有可选特征.节点基尼指数描述了节点的不纯度,计算公式如下:

$$Gini(t)=1-\sum_{j=1}^{k}[p(j\mid t)]^2 \tag{1}$$

其中,$p(j|t)$表示样本点在节点 t 处属于 j 类的概率.基尼指数越大,表明在节点 t 处的样本数据越均匀,所含信息就越少.

4) 每棵决策树都最大可能地进行生长而不进行剪枝.

每棵决策树由节点和有向边组成,节点有两种类型:内部节点表示一个特征或属性,叶节点表示一个类别.图 1 展示了决策树的结构,其中,$A1$,$A2$ 是内部节点,表示特征或属性;$C1$~$C3$ 是叶节点,表示类别.

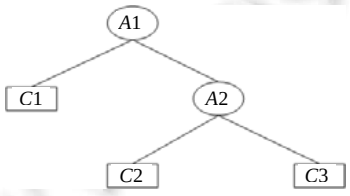

Fig.1 Decision tree structure

图 1 决策树的结构

2.2 随机森林投票

随机森林在面对分类问题时,一般采取的是简单投票法.测试数据输入到每个基决策树中进行分类,最终的

类别由各个基决策树的分类情况决定,取分类得票数最多的那一类作为最终结果.即对于测试数据 X,每棵决策树预测该数据的类别为 C,则随机森林的投票决策公式如下:

$$C_p = \arg\max_c \left(\frac{1}{n_t} \sum_{i=1}^{n_t} I\left(\frac{n_{s_i,c}}{n_{s_i}} \right) \right) \tag{2}$$

其中,n_t 表示基决策树的总数,$I(*)$表示示性函数,$n_{s_i,c}$ 表示类别 C 在树 s_i 上的分类结果,n_{s_i} 表示叶节点个数.

3 基于维修日志的故障诊断基本过程

基于维修日志的故障诊断的核心思想是:通过机器学习方法,利用飞机历史维修经验,为新的故障诊断提供依据和参考.诊断过程如图 2 所示,包括故障数据转换、故障原因判断、故障原因修正和故障案例添加等步骤.

1) 故障数据转换:由于故障案例包含故障现象、故障位置等均为维修人员用自然语言书写的文字记录,无法直接计算,故采用词频-逆文本频率、独热编码等方法将非结构化的文本转化为结构化数据,形成待预测故障.
2) 故障原因判断:通过随机森林等机器学习模型对测试故障样本的故障原因进行诊断,并显示给维修人员作为故障诊断的参考依据.
3) 故障原因修正:通过专家的经验对测试故障样本的故障原因进行修正,对预测错误的故障原因进行修正,形成正确的故障原因,以作为新的数据训练故障诊断模型,提高预测的准确率.
4) 故障案例添加:将修正的故障样本添加到已有的故障库中,每隔一定时间,迭代地训练新的故障库,以提高故障诊断模型的预测准确率.

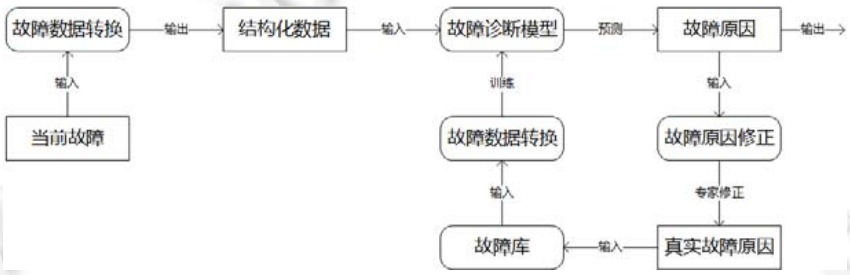

Fig.2 Fault diagnosis process based on maintenance log

图 2 基于维修日志的故障诊断过程

4 文本特征提取方法

航空维修数据一般为模块化的非结构化的文字记录,维修人员观察飞机设备的故障现象等故障信息之后,用专门的测试设备对疑似故障部件进行测试,根据维修经验进行故障诊断并记录在故障数据库中.本文采用的航空维修数据是针对波音 737-300 近 7 年的故障诊断记录,该数据来自于某合作单位,数据的样例见表 1.

Table 1 Maintenance log of Boeing 737-300

表 1 波音 737-300 维修日志

故障现象	故障失常码	故障件位置	系统	工作时次	故障原因
襟翼放 5 度,1 号缝翼半伸出绿灯不亮	工作不正常、失常	机翼	自动驾驶设备	3 500	传感器故障
客舱第 1 包间灯管不亮	灯不亮	客舱	电气装置	5 100	灯组件故障
刹车压力指示器压力偏大	工作不正常、失常	起落架舱	信号系统	9 060	指示器故障
自动油门出现卡滞	卡滞(紧涩)	NULL	自动驾驶设备	5 500	机械故障
左侧航行灯灯罩有裂纹	断裂、破裂、折断	NULL	供电系统	10 020	设备老化

该维修数据的维修机型均为波音 737-300,一共统计了 3 架飞机的维修情况,但只有 1 架飞机拥有 2010 年~2016 年近 7 年的维修数据,另外两架只包含部分年份的维修数据.在数据库中,维修人员记录故障现象、故

障失常码、故障件位置、故障所属系统、工作时次和故障原因.其中,故障现象为故障发生时维修人员看到的故障情况,并通过文字记录的形式存入到数据库中.由于维修人员的更替,这种文字记录形式不统一,不同的维修人员对同一故障现象的记录可能会有差异.故障失常码为故障发生时对故障表现的简要总结,包括工作不正常、灯不亮、不指示、噪音大等.故障件位置是故障发生时故障设备所处飞机的位置,有前机身、前设备舱、机翼、客舱等等.系统表示了该种故障现象发生在飞机的哪种系统中,如自动驾驶设备、电气装置、信号系统等等.工作时次表示了故障部件已经工作的时长,以小时为单位.故障原因为本文方法预测的目标,主要包括传感器故障、灯组件故障、电路故障、电门故障、机件内部故障等 11 个故障原因.

4.1 基于卷积神经网络的文本特征提取方法

故障现象作为维修人员观察故障特征的主要记录手段,揭示了故障表现与故障原因的内在关系.但维修记录中的文字描述缺乏统一的描述规范,同一故障现象的描述方式可能有所不同.并且由于维修人员不断更换,文字记录的方式往往伴随着随机性.设计一种能够从自然语言文本中提取核心特征的方法,是做故障诊断任务之前的关键.文献[20]在预训练的词向量上直接使用一个简单的卷积网络用于句子级别的分类任务中,并在4种领域问题如情感分析、问题分类等做了验证,证明卷积神经网络能够较好地提取文本的特征.文献[21]使用英文字符为单位的卷积网络实现文本分类,在与传统模型和深度学习模型进行比较实验的过程中,表明了字符级卷积网络可以获得具有竞争力的结果,但该种方法的缺陷在于需要大量的语料库的支持才能获得较好的效果.文献[22]在语义匹配领域中提出了一种不需要先验知识的卷积神经网络模型,通过使用卷积来代表两个句子的层次结构并捕获丰富的匹配模式,可以应用于不同性质和不同语言的匹配任务中,通过实验证明了对各种匹配任务的有效性及相对于其他模型的优越性.故本文采用卷积神经网络对该类文本进行结构化转换.

卷积神经网络主要结构如图 3 所示,该结构将“故障原因”作为目标进行训练,以字向量为输入方式,最终通过全连接层间接得到故障现象中维度固定的文本向量.下面主要讲解神经网络各层的作用.

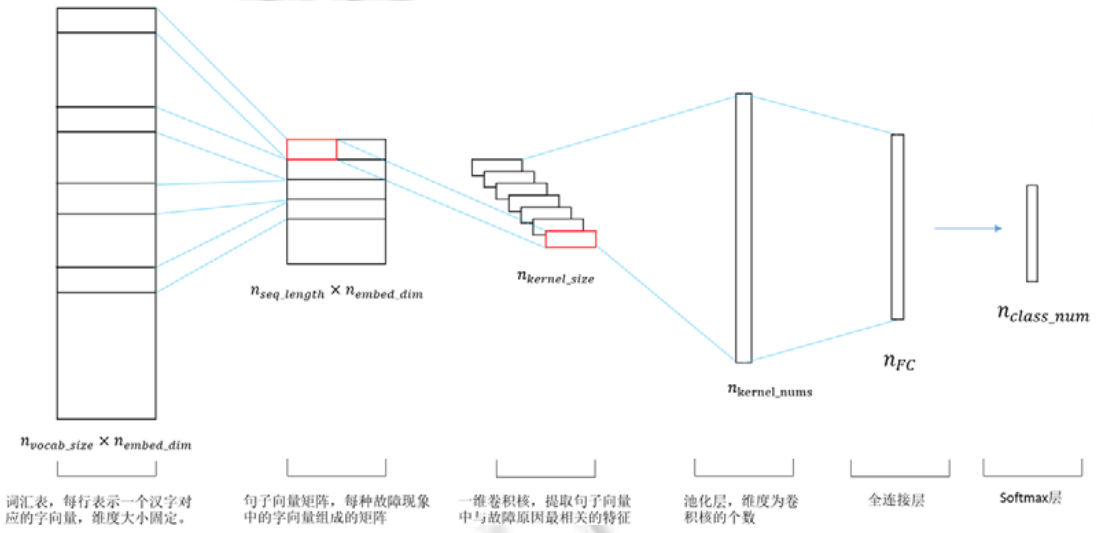

Fig.3 Convolutional neural network structure

图 3 卷积神经网络结构

1. 统计故障现象中出现过的所有文字并建立词汇表.故障现象中包含大量航空领域中的领域故障词,将词汇表中的文字依出现次数进行排序,并为每个文字分配一个序号,出现次数更多的文字,其排序更为靠前(添加特殊字符〈UNK〉作为未出现在该词汇表中的字).为排序在前 n_{vocab_size} 位的词汇建立大小为 $n_{vocab\_size} \times n_{embed\_dim}$ 的词汇表,并随机初始化我们的输入——字向量.其中,n_{embed_dim} 为字向量的维度.
2. 由于故障现象描述中文字的数量不一致,为了保持统一,选择一个合适的大小 n_{seq_length} 作为该句中需

要提取的字向量个数,从而构建句子向量矩阵.如果句子中的字数较少,则补全为空;如果句子中的字数较多,直接截断前 n_{seq_length} 个字.

3. 选择一维卷积核提取句子特征,核大小为 n_{kernel_size},选择 n_{kernel_nums} 个卷积核构建卷积层,对句子向量矩阵做卷积运算.卷积运算是将核权重 $\omega \in \mathbb{R}^{1\times n_{kernel\_size}}$ 与窗口大小为 n_{kernel_size} 的字向量相乘,并得到新的特征,计算公式如下:

$$c_i = f(w \cdot x_{i:i+n_{kernel\_size}-1} + b) \tag{3}$$

4. 使用最大池化层提取每行的最大值作为该卷积核提取出来的特征,形成 n_{kernel_nums} 大小的池化层,并与全连接层相连,其中,使用 dropout 随机失活等正则化方式防止过拟合.该全连接层代表着整个句子经过卷积核的特征提取后的向量表示.
5. 将全连接层与 *softmax* 层相连,*softmax* 层的维度为故障原因类别个数,将属于某一类的故障原因的索引设置为 1,其他设置为 0.
6. 输入故障现象和故障原因,训练整个神经网络,得到更能反映领域知识的文本向量.

在采用合适的正则化策略与激活函数之后,该类方法能够在全连接层提取故障现象中最能反映故障原因的文本特征,从而将该特征作为故障现象的文本特征与其他领域特征进行拼接,得到合适的结构化文本.该类方法不仅可以提取到表示层次较深的文本特征,而且可以降低文本表示的维度,得到紧凑稠密的文本表示.

4.2 基于独热编码的文本转换方法

独热编码(one-hot encoding)又称为一位有效编码,它使用 *N* 位向量表达 *N* 个词是否出现,第 *i* 个位置为 1 表示第 *i* 个词在文本中出现.从计算机体系结构角度来看,其实对 *N* 个不变状态采用 *N* 位寄存器来保存,每个寄存器只保存 1 种状态,并且在任意时刻只有 1 个寄存器有对应的状态.

其他特征取值范围固定,所以采用独热编码的方式将每一行的相应特征转换为向量的格式用于计算.如故障件位置包括后机身、机翼、客舱、起落架舱等 7 个位置,加上记录为空的字段,转换为独热编码即为 8 维向量.将故障失常码、故障件位置、系统这 3 个特征均做独热编码处理.

“工作时次”表示设备到故障为止的正常运行的小时数,原数据为浮点数格式,为了防止过拟合,将其分为 10 个子范围,并采用独热编码转换为向量的格式用于计算.

4.3 基于随机森林的故障诊断步骤

图 4 展示了基于维修日志数据的基于随机森林算法的故障诊断过程.

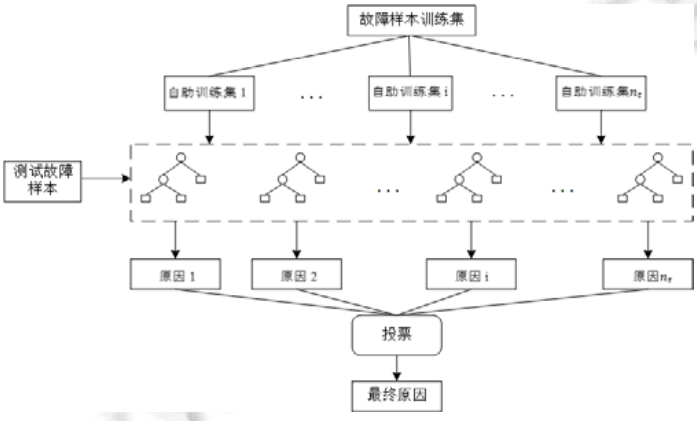

Fig.4 Fault diagnosis process based on random forest

图 4 基于随机森林的故障诊断流程

整个过程的步骤如下.

1) 获取经过文本处理后的原始故障样本训练集$(x_i,y_i)_{N\times M}$,x_i表示第 i 个故障样本的特征向量,y_i表示该故障样本的真实故障原因,N表示故障样本数,M表示特征数.
2) 将原始故障样本训练集分为 n_t 个自助训练集,根据上述的随机森林构建方法构建 n_t 棵基决策树.
3) 将测试故障样本输入到构建的随机森林模型中,每棵基决策树分别判断该故障样本的故障原因.
4) 利用投票法综合考虑所有基决策树的分类结果,由公式(3)得出该故障样本的故障原因.

5 实验设计与结果

5.1 数据集

实验的数据集来自于真实的波音737-300飞机维修日志,该数据集记录了近7年的故障诊断记录,包括飞机编号、故障发现日期、故障现象、系统、工作时次、故障失常码和故障原因等信息.其中,飞机编号在本实验中没有作用,与故障原因关系不大,故在实验中删除该列.故障发现日期将作为故障样本训练的基准,迭代地训练诊断模型.原始数据集中,故障原因中有些记录为冗余信息,有些记录不够明确,仅仅通过原始故障原因无法有效完成故障诊断模型的建立,因此通过对故障原因类别的梳理,对其记录中的主要信息提取整理,最终得到处理后的故障原因,见表2.数据集共有故障样本1 272个,而故障原因作为预测的目标,其样本数分布见表3.

Table 2 Partial aircraft failure raw data

表 2 部分飞机故障原始数据

原始故障原因	预处理后的故障原因
燃油交输活门电插头老化	设备老化
大气数据计算机内部控制模块失效	计算机故障
振动指示器内部摩擦大	机械故障
压力传感器线圈阻值低	电阻故障

Table 3 Number of samples for each failure reason

表 3 各故障原因样本数

故障原因	样本数
传感器故障	87
指示器故障	99
机件内部故障	223
机械故障	134
灯组件故障	99
电路故障	150
电门故障	105
电阻故障	84
计算机故障	97
设备烧蚀	90
设备老化	104

5.2 卷积神经网络参数调整

卷积神经网络虽然能够较好地提取领域内的字向量特征,但与其他文本特征提取方法相比,模型的复杂程度更高,需要调节的参数也变得更多.本文中采用的参数调节方式主要使用训练集中的故障现象去尽可能得到最佳的故障原因预测精度,参数调节过程中定义的参数搜索域见表4.

Table 4 Convolutional neural network parameter search domain

表 4 卷积神经网络参数搜索域

字向量维度	句向量字数	卷积核数	卷积核大小	词汇表大小	全连接层大小	Dropout 比例	学习率	batch_size
10	64	128	5	600	128	0.9	1e-4	64
20	48	64	6	600	96	0.8	1e-3	32
30	32	32	7	600	64	0.5	1e-2	16

参数调节过程的评价指标为准确率,经过若干次随机搜索,取达到最高准确率的超参数作为模型的超参数来使用,并使用该模型得到的故障现象文本表示与其他领域特征结合并用于随机森林模型中.

5.3 评价指标与模型参数

分类问题中常用的评价指标是准确率和召回率,除此之外,本文还采用了 $F1$ 值作为综合考虑准确率和召回率的评价指标.上述指标均是数值越大,表示模型效果越好.

为了让随机森林算法在该数据集上达到最好的效果,需要调整算法的超参数使其更适合该类数据.随机森林算法主要包括两个参数:随机选择的特征数 m 和基决策树数目 n_t.

随机选择的特征数 m 为每棵树的节点在进行分裂时需要考虑的特征数量,它是随机森林算法中对准确率预测比较重要的参数.调整 m 的取值,随机森林的性能会随之变化.本文通过实验来确定最佳的特征数 m:首先固定基决策树个数 n_t 为 100,调整 m 的取值,观察随机森林在该数据集上的 $F1$ 值变化,选择 $F1$ 值最大时的 m 值作为本文实验中 m 的取值.图 5 展示了随机森林与不同 m 值之间的关系,由于数据维度较高,故 m 值代表取原始特征数的比例,取值范围为 0.1~1.0.由图 5 可知,当 m 值为 0.1 时效果最好.

随机森林是由许多基决策树组成,基决策树的数量与随机森林的预测性能有较大的关系.基决策树数量足够多,随机森林才能达到更高的误差上界.但若基决策树数量过多,随机森林的训练时间也会变长且容易造成过拟合,在测试数据集上表现不佳.本文通过实验确定最佳的 n_t,首先固定 m 值为 0.1,选择不同的 n_t 在数据集上进行训练,调整 n_t 的取值观察随机森林在该数据集上的 $F1$ 值变化,选择 $F1$ 值最大时的 n_t 值作为本文实验的取值.图 6 展示了随机森林与不同 n_t 值之间的关系,取值范围 n_t=[20,50,100,150,200,300,500,750,1000].由图 6 可知,当 n_t 为 500 时效果最好.

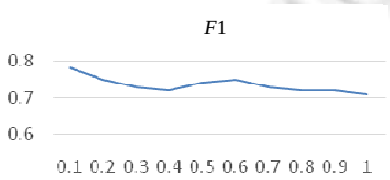

Fig.5 $F1$ value corresponding to different m values

图 5 不同 m 值对应的 $F1$ 值

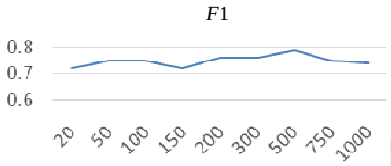

Fig.6 $F1$ value corresponding to different n_t values

图 6 不同 n_t 值对应的 $F1$ 值

5.4 实验结果分析

文本的实验分为 3 部分:第 1 部分使用上述讨论的超参数对故障训练样本进行迭代式地学习,将次年的故障样本作为测试集,逐年添加故障样本,观察随机森林模型在测试集上的预测能力;第 2 部分将随机森林模型与其他机器学习算法在该数据集上作对比,观察随机森林相比于其他算法的优越性;第 3 部分对比不同的文本特征提取方式在同一模型下对模型准确程度的影响.

首先,观察随机森林模型在迭代式地学习中获得的提升幅度.把故障数据逐年递增式地输入到超参数已定的随机森林算法中,测试集选择次年的数据,比如 2010 年~2012 年的数据作为训练集,则 2013 年的故障样本便作为测试集,观察平均准确率、平均召回率和平均 $F1$ 值的变化.最终的实验结果见表 5.

Table 5 Iterative training classification result

表 5 迭代训练的分类结果

故障样本年份	平均准确率	平均召回率	平均 $F1$ 值
2010~2011	0.66	0.60	0.63
2010~2012	0.69	0.67	0.68
2010~2013	0.73	0.68	0.70
2010~2014	0.76	0.74	0.75
2010~2015	0.83	0.82	0.82

由表 5 中我们可以明显观察到,随着故障样本迭代式地增多,3 个模型评价指标都随之增长,在故障年份为

2010 年~2015 年时,3 个模型评价指标均为最高,达到了 82%左右.由此可以证明,随着故障库中故障样本的增加,模型的预测能力确实有了显著的提高.

现把 2010 年~2015 年的故障数据全部输入到超参数已定的随机森林算法中,预测 2016 年的故障样本的故障原因,观察各个故障原因在该算法下的分类性能.最终的实验结果见表 6.

Table 6 Random forest classification result

表 6 随机森林分类结果

故障原因	准确率	召回率	*F*1 值
传感器故障	0.87	0.87	0.87
指示器故障	0.86	0.89	0.87
机件内部故障	0.78	0.85	0.81
机械故障	0.84	0.76	0.80
灯组件故障	0.75	0.72	0.73
电路故障	0.68	0.72	0.70
电门故障	0.81	0.84	0.82
电阻故障	0.85	0.92	0.88
计算机故障	0.91	0.88	0.89
设备烧蚀	0.90	0.97	0.93
设备老化	0.88	0.62	0.72
平均值	0.83	0.82	0.82

在表 6 中我们观察到,在准确率方面,电路故障最低,其他故障的预测准确率都在 70%以上,其中,计算机故障和设备烧蚀故障的分类准确率最高,都超过了 90%;在召回率方面,设备老化最低,没有超过 70%,而电阻故障和设备烧蚀故障的召回率均超过了 90%;*F*1 值在一定程度上反映了学习器在准确率和召回率上取得双高的比例,电路故障、灯组件故障和设备老化故障的 *F*1 值最低,没有超过 80%,*F*1 值最大的为设备烧蚀故障,达到了 93%.可以看到,在对维修日志数据的分类预测问题上,随机森林有较好的分类效果.

在不同模型的对比实验方面,我们采用逻辑回归、朴素贝叶斯、决策树、支持向量机和 *k* 近邻算法与随机森林算法进行对比.其中,逻辑回归使用“l2”正则化,朴素贝叶斯使用多项式模型,决策树使用 CART 决策树,支持向量机核函数使用高斯核函数,*k* 近邻的距离度量方式选择欧氏距离.经过参数调整以后,各个模型的最佳实验结果见表 7.

Table 7 Comparative experimental results of different algorithms

表 7 不同算法的对比实验结果

算法名称	平均准确率	平均召回率	平均 *F*1 值
逻辑回归	0.69	0.67	0.68
朴素贝叶斯	0.68	0.59	0.63
决策树	0.76	0.74	0.75
k 近邻	0.53	0.49	0.51
支持向量机	0.73	0.70	0.71
随机森林	0.83	0.82	0.82

由表 7 可知,其他 5 种算法的预测性能均没有随机森林强,其中,*k* 近邻的效果最差.这是因为除了字特征可能具有可以衡量的距离以外,其他特征并不具有明显的距离概念,并且在高维数据下采用欧式距离可能达不到很好的度量效果.而逻辑回归和朴素贝叶斯的效果都低于 70%,在该类数据上的效果比较差.决策树的效果要稍好一些,可能是因为决策树模型与传统上专家在进行故障诊断时所依据的故障树规则比较相似,所以达到了比较好的效果.而支持向量机在面对小样本集时也能发挥其良好泛化性能的特点.随机森林结合了决策树模型的优点,同时通过 Bagging 集成的方式降低了算法的泛化误差,获得了最优的效果.

最后比较不同文本特征提取方式对模型性能的影响,分别采用直接独热编码的词袋模型、TF-IDF、基于维基百科语料训练的分布式词向量、基于故障现象小样本语料训练的分布式词向量和本文提出的字向量特征,并使用同一参数的随机森林模型进行预测,观察效果.实验结果见表 8.从结果中可以看到,直接使用词袋模型的

独热编码方式由于无法提取文本特征的上下文特征和领域特征,效果最差;而 TF-IDF,Word2Vec+维基百科语料的准确度比较接近;而 Word2Vec+故障现象语料的准确度也较差,这是因为故障现象语料较少,直接使用上下文关系预测词向量的方式效果较差;而采用卷积神经网络提取基于字符级的字向量特征的方法比其他方法在总体性能上更好,平均各项指标比 Word2Vec+维基百科语料要高 0.03,从而可以说明采用卷积神经网络的文本特征提取方式对文本特征提取更加充分,更能反映与故障原因的关系.

Table 8 Comparative experimental results of different text features

表 8 不同文本特征的对比实验结果

特征提取方法	平均准确率	平均召回率	平均 *F*1 值
词袋模型	0.75	0.73	0.74
TF-IDF	0.79	0.77	0.78
Word2Vec+维基百科	0.80	0.78	0.79
Word2Vec+故障现象	0.78	0.75	0.76
卷积神经网络	0.83	0.82	0.82

6 结 论

针对目前维修日志数据无法充分利用的问题,本文首先提出一种迭代式的故障诊断基本过程,然后提出一种基于卷积神经网络对非结构化文本使用字向量提取文本特征的方法,最后使用随机森林算法对长期积累下来的飞机故障日志数据建立故障原因分类器,并通过实验验证了文本特征提取方式和随机森林算法的有效性.并且如果后续有更多的故障日志数据作为支撑,可直接使用本文阐述的故障诊断过程,通过迭代的方式不断地提高故障诊断模型的预测精度,帮助维修人员尽快确定故障原因,节省维修人员的时间成本.但是随着维修日志的不断增多,维修日志语料库也在不断增多,有必要继续比较本文提出的文本特征提取方式与基于上下文关系的文本特征提取方式的优劣.同时,有必要解决样本量较少且样本不均衡等问题,这是我们下一步的研究方向.

References:

[1] Kramer MA, Palowitch Jr BL. A rule-based approach to fault diagnosis using the signed directed graph. AIChE Journal, 1987,33(7): 1067−1078.

[2] Lin CE, Ling JM, Huang CL. An expert system for transformer fault diagnosis using dissolved gas analysis. IEEE Trans. on Power Delivery, 1993,8(1):231−238.

[3] Gertler J. Fault Detection and Diagnosis. London: Springer-Verlag, 2013.

[4] Zhang J. Improved on-line process fault diagnosis through information fusion in multiple neural networks. Computers & Chemical Engineering, 2006,30(3):558−571.

[5] Thukaram D, Khincha HP, Vijaynarasimha HP. Artificial neural network and support vector machine approach for locating faults in radial distribution systems. IEEE Trans. on Power Delivery, 2005,20(2):710−721.

[6] Jing C, Hou J. SVM and PCA based fault classification approaches for complicated industrial process. Neurocomputing, 2015,167: 636−642.

[7] Yan W. Application of random forest to aircraft engine fault diagnosis. In: Proc. of the IMACS Multi Conf. on Computational Engineering in Systems Applications. IEEE, 2006. 468−475.

[8] Chiu C, Chiu NH, Hsu CI. Intelligent aircraft maintenance support system using genetic algorithms and case-based reasoning. The Int'l Journal of Advanced Manufacturing Technology, 2004,24(5-6):440−446.

[9] Li Q, Shi YQ, Zhou Y. CBR methodology application in fault diagnosis of aircraft. Journal of Beijing University of Aeronautics & Astronautics, 2007,33(5):622−626 (in Chinese with English abstract).

[10] Zhao Y, Xu TH. Text mining based fault diagnosis for vehicle on-board equipment of high speed railway signal system. Journal of the China Railway Society, 2015,37(8):53−59 (in Chinese with English abstract).

[11] Rajpathak DG. An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain. Computers in Industry, 2013,64(5):565−580.

[12] Wang F, Xu T, Tang T, *et al*. Bilevel feature extraction-based text mining for fault diagnosis of railway systems. IEEE Trans. on Intelligent Transportation Systems, 2017,18(1):49−58.

[13] Zhao Y, Xu T, Hai-Feng W. Text mining based fault diagnosis of vehicle on-board equipment for high speed railway. In: Proc. of the 2014 IEEE 17th Int'l Conf. on Intelligent Transportation Systems (ITSC). IEEE, 2014. 900−905.

[14] Ramos J. Using TF-IDF to determine word relevance in document queries. In: Proc. of the 1st Instructional Conf. on Machine Learning, Vol.242. 2003. 133−142.

[15] Liaw A, Wiener M. Classification and regression by random forest. R News, 2002,2(3):18−22.

[16] Li C, Sanchez RV, Zurita G, *et al*. Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals. Mechanical Systems and Signal Processing, 2016,76:283−293.

[17] Cerrada M, Zurita G, Cabrera D, *et al*. Fault diagnosis in spur gears based on genetic algorithm and random forest. Mechanical Systems and Signal Processing, 2016,70:87−103.

[18] Santur Y, Karaköse M, Akin E. Random forest based diagnosis approach for rail fault inspection in railways. In: Proc. of the 2016 National Conf. on Electrical, Electronics and Biomedical Engineering (ELECO). IEEE, 2016. 745−750.

[19] Breiman L. Random forests. Machine Learning, 2001,45(1):5−32.

[20] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[21] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Proc. of the Advances in Neural Information Processing Systems. 2015. 649−657.

[22] Hu B, Lu Z, Li H, *et al*. Convolutional neural network architectures for matching natural language sentences. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 2042−2050.

附中文参考文献:

[9] 李青,史雅琴,周扬.基于案例推理方法在飞机故障诊断中的应用.北京航空航天大学学报,2007,33(5):622−626.

[10] 赵阳,徐田华.基于文本挖掘的高铁信号系统车载设备故障诊断.铁道学报,2015,37(8):53−59.

王锐光(1993－),男,河北沧州人,硕士,CCF 学生会员,主要研究领域为安全关键系统与软件,智能化软件测试.

刘超(1958－),男,博士,教授,CCF 高级会员,主要研究领域为软件工程,软件质量,软件测试,模型驱动软件开发方法.

吴际(1974－),男,博士,副教授,CCF 专业会员,主要研究领域为安全关键系统与软件,智能化软件测试.

杨海燕(1974－),女,讲师,主要研究领域为安全关键系统与软件,智能化软件测试.