基于机器学习的空管安全信息处理方法初探

鲁翰涛，郑灶旭，黄跃鸿
（中国民用航空中南地区空中交通管理局，广东 广州 514000）

摘　要：空管系统安全信息的处理和挖掘多来自报告人对信息的模块化录入，不仅耗费人力物力，而且容易因怕麻烦而发生漏报和瞒报行为。将中文分词以及决策树、随机森林算法等机器学习技术应用于空管不安全事件信息的处理，对中南空管系统 2015~2019 年的典型不安全事件信息进行机器识别，准确率达到 71%，有效降低了信息报送的人力和时间成本，进一步提升了安全管理部门的信息处理效率。

关键词：空管安全信息；机器学习；安全信息处理；决策树；随机森林
中图分类号：X913.2　　**文献标识码：**B

Research on the Method of ATC Safety Information Processing Based on Machine Learning

LU Hantao, ZHENG Zaoxu, HUANG Yuehong
(Central South Air Traffic Management Bureau CAAC, Guangzhou 514000, Guamgdong, China)

Abstract: The processing and mining of air traffic control system safety information mostly come from the reporter's modular inputting, which not only consumes human and material resources, but also is prone to under-reporting due to fear of trouble. Machine learning technologies such as Text statistical segmentation, decision tree based on key words and random forest algorithm are applied in the processing of air traffic controlsafety information . The machine recognition of typical safety information of the Contrae South ATM Buredu during 2015-2019 is carried and reaches with the accuracy of 71%, which greatly reduces the labor and time cost of information submission and further improves the information processing efficiency of safety management department.

Key words: ATC safety information; machine learning; safety information processing; decision tree; random forest

0　引言

国际民航组织在国际民航公约的第 19 个附件（安全管理）和 DOC9859（安全管理手册）中提出：各国必须建立安全数据收集和处理系统，以便收集、存储、汇总和分析安全数据和安全信息。安全数据（最初报告或记录）在特定的环境下经过处理、组织、综合或分析后，便转换为安全信息。安全信息可以继续以不同的方式予以处理，以提取不同的含义，使其有助于安全管理。中国民航和空管行业内部均建立了强制性和自愿性安全信息报告机制。按照《民航空管系统安全信息管理办法》要求，空管不安全事件信息需要在规定时限内上报包含事发地点、事件原因、事件类型和事件经过等 20 多项要素的安全信息报告表。中南空

作者简介：鲁翰涛，男，湖南常德人，工程师，主要从事民航空中交通管制、空管安全管理等研究。

管局开发了 SIRAMS 系统（中南空管安全信息及运行危害预警管理系统）辅助信息报告人填选相关要素，但在初始报告阶段，安全信息的录入依然耗费不少时间，人为错误数量也比较高。

此外，空管系统安全管理人员对不安全事件信息的处理多依靠人工审核、分类，以及基于已知类别的数据统计。由于人的主观理解和判断偏差，以及现有类别的局限性，安全信息质量和数据分析的深度还存在不足。例如对偏离空管指令高度的事件原因分析，大多能分析到管制、天气、航空器等原因的偏离事件数，却无法统计到更细、更深层原因（例如：高度指令错误、机组复诵错误或管制监控不到位等）的事件数，主要困难在于人工统计不能处理大量数据，机器计算却缺少原始数据的深层次原因分类，数据颗粒度不够。如果需要精准感知安全态势和提供决策参考，对事件内容的精细化、深层次分析将是未来安全信息处理的主要需求。

1　国内外现状

对于不安全事件信息的文本识别和机器处理，国内多个专业院校的学者进行过测试性的研究，主要有：张聪俊 [1] 对冲 / 偏出跑道等不安全事件报告的机器分类关键方法进行了研究；田继存 [2] 对民航安全自愿报告系统中的数据进行了文本分类研究；崔振新、卢昊文 [3] 对民航安全信息关键词提取方法进行了研究等。这些研究都需要对文本进行分词，因缺少全面的空管安全信息词典支撑而影响了分类的准确率。

国外民航组织和机构对安全信息的深度处理和统计分析已初具规模。比如：国际航协（IATA）通过 GADM (Global Aviation Data Management) 开发了交互式基准测试和查询工具。美国的 MITRE 公司开发了 Aviation Safety Data Mining Workbench，用于对美国航空公司 ASAP 数据 (American Airline's Aviation Safety Action Program data) 进行文本挖掘和数据处理分析。欧洲民航安全局（EASA）则通过 NoA (Network of Analysts) 和 CAGs (Collaborative Analysis Groups) 定期进行系统性的安全分析和专业安全信息的集中处理。这些系统多基于安全信息初始报告的筛选和分类，着重于事后的数据挖掘和统计分析，少有对初始报告的自动识别、分类或聚类处理。

2　技术理论综述

空管安全信息是一种中文为主、专业术语和自然语言高度融合的文本。如要实现自动处理，需要以下两方面技术。

2.1　安全信息文本分词技术

安全信息文本分词是安全信息文本处理的基础技术。分词的准确率和有效性对安全信息处理的准确性起着至关重要的作用。目前，中文分词技术可分为 3 类：基于词库匹配的分词技术、基于规则和语义的分词技术、基于统计理论的分词技术。这些方法优缺点各异：基于词库匹配的分词技术快速直接，但对词汇的直接匹配容易出现歧义或割裂的情况；基于规则和语义的分词技术需要高度概括、总结中文语法句式来满足计算机的模拟需求，但中文的博大精深使这种方法还处在探索阶段；基于统计的分词技术指对已知词库中词汇和多歌词搭配关系出现的频率进行统计，其分词性能高于其他方法，但该技术对已知词库的预处理有较高要求，同时需要大规模数据频率计算，实时性与速度稍略逊于另外 2 种方法。

2.2　安全信息文本分类技术

安全信息文本分类指的是机器对安全信息中的事件经过文本按照既定分类规则或标准完成的自动分类，主要过程包括以下步骤。

1）文本格式化：统一事件经过文本格式，整合分类标签，为分词处理做好准备。

2）确定文本特征：梳理分词描述数据，选取符合分类标签或主题的特征并进行概率计算。

3）设计分类模型：选取数学模型表达文本特征，优选决策树和随机森林算法构建模型。

4）学习和验证：将标注过的安全信息文本输入分类模型进行监督学习、不断调参和测试，然后用验证数据验证分类模型的准确率，达到所需性能。

3　安全信息的机器处理方法研究

安全信息机器处理是空管安全管理高质量发展的重要标志。其处理方法主要包括安全信息词典的构建、文本向量化、分类模型设计、验证和测试几个步骤，能够有效依托现有安全管理系统为用户提供方便、快速的安全信息报告录入，提高安全信息处理能力和效率。

3.1　安全信息词典搭建

目前，国内空管行业的词典尚属空白，个别企业（比如蓝天航空词典和云帆民航词典）的民航行业词典中，空管专业的词条仅千余条，在航行情报数据方面比

较欠缺，难以满足空管安全信息分词的需求。本文通过输入空管专业文档、资料和航行情报数据，基于 Python 开发的 Jieba 分词算法，经过词图扫描、查找基于词频的最大切分组合等方法，结合基于汉字成词能力的 HMM 模型和 Viterbi 算法发现新词，然后过滤停用词和通用词，得到了一个包含 2 万多词条的词典原型，具体经过如下。

1）基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG）。Trie 树是有名的前缀树，如果几个词语的前面几个字一样，就表示他们具有相同的前缀，可以使用 Trie 树来存储，具有查找速度快的特点。对待分词句子，由于 Jieba 算法自带一个 dict.txt 的词典，包含 2 万多条词，分词函数为：cut(words,HMM=True)，可根据 dict.txt 生成的 Trie 树构成有向无环图。

2）用动态规划查找最大概率路径，找出基于词频的最大切分组合。首先查找待分词句子中已经切分好的词语，再查找该词语出现的频率（次数 / 总数）。然后根据动态规划查找最大概率路径的方法，对句子从右往左反向计算最大概率，得到最大概率的切分组合。

3）对于未登陆的新词，采用 HMM 模型和 Viterbi 算法对大量语料进行训练，得到多个概率表，进而产生概率最大的的 BEMS 序列，按照 B 打头，E 结尾的方式，对待分词的句子重新组合，得到分词结果。

4）用通用的停用词库和百度词库对分词结果进出过滤，去除包括语气助词、副词、介词、连接词和通用领域的常见词语、数字词语等，得到了 20 750 个词语。根据词语出现的次数进行排序，保存为 txt 文件形成分词词典。图 1 是出现频次最高的部分词语。

管制运行部进近
管制室塔台管制
塔台管制室安全管理
管制席进程单
管制单位安全信息
管制员质量安全
应急处置进近管制室
空管站

图 1 分词词典中的高频词语

3.2 文本向量化

机器学习算法基于数学计算，输入的文本信息要先转化成数字向量，所以必须对安全信息的分词结果进行向量化。本文使用基于 TF-IDF（term frequency - inverse document frequency）的词袋模型，该模型是在自然语言处理下的表达模型[4]。在此模型下，一段文本可以用一个装着这些词的袋子来表示，该方式不考虑文法以及词的顺序，只将每个文本词汇在词袋中出现的次数组成向量。例如，有一段文本为“停机位被占用，等待飞机离开停机位”，若将词袋候选词设置为（“停机位”、“占用”、“等待”、“飞机”、“离开”），则这段文本对应的词袋向量为（2,1,1,1,1），另一段文本“飞机遭遇鸟击”对应的向量为（0,0,0,1,0）。

词袋候选词不包含所有词汇，而是选择有代表性的特征词汇。TF-IDF 是一种统计方法，用于评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，随着它在语料库中出现的频率成反比下降。一个词汇在文件中的重要程度如式（1）。

$$tfidf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}\lg\frac{D}{c_{i,j}} \tag{1}$$

式（1）中，$n_{i,j}$ 表示词 t_i 在文件 d_j 中出现的次数，$\sum_k n_{k,j}$ 表示文件中所有字词的出现次数之和，D 表示语料库中文件总数，表示包含词语 t_i 的文件数目。

本文使用 TF-IDF 算法评估词汇对文本集的重要程度，在经过数据清洗后的 17 个事件类型样本中，为每个数据类型选择 TF-IDF 值最大的 30 个词作为特征词，将这些特征词合并后去重，作为词袋候选词集合。

在后续的训练中，词汇的权重与词汇所处的文件无关，词的权重取平均值，词 t_i 的权重的计算如式（2）所示。

$$w_i=\frac{\sum_{j=1}^{n} tfidf_{i,j}}{c_i} \tag{2}$$

式（2）中，n 为包含词的文件数量。

3.3 分类模型构建

（1）决策树算法

决策树训练是数据挖掘中的常用方法，目标是创建一个模型预测样本的目标值[5]。本文的数据类标是事件类型，属于多分类任务，使用 python 语言进行编程，决策树分类器已被封装在 Scikit learn 库的 DecisionTreeClassifier。

（2）随机森林算法

随机森林是利用多棵树对样本进行训练并预测的一种分类器，通过多次随机选取部分样本和特征构建多颗决策树，其输出的类别由个别树输出的类别的众数而定[6]。相对于决策树，它具有更强的准确率和鲁棒性。本文使用 Scikit learn 库中封装好的 RandomForestClassifier 构建随机森林模型。

3.4　算法验证和模型优化

搜集中南空管系统 2015～2019 年的空管不安全事件信息作为数据样本，共计安全信息 18 478 条，均标注有事件类型、原因和结果等。选取空中颠簸、起落架故障、无线电干扰等 17 个不安全事件类型测试不同的分类算法。每次均用 80% 的事件经过数据作为训练样本，其余 20% 作为测试样本。

在进行模型训练之前，选择准确率作为模型的评价指标，准确率为预测正确的样本数除以总样本数。

3.4.1　算法验证

为了验证不同的特征构建和模型构建方法在此任务中的表现，本文设计了一个实验：分别将决策树和随机森林算法与 4.2 中的词袋模型组合，得到以下 2 种算法。

算法 1：TF–IDF+ 决策树；算法 2：TF–IDF + 随机森林。

以上决策树和随机森林均使用默认参数进行实验，得到实验结果如下：算法 1 的准确率是 59.5%；算法 2 的准确率是 65.8%。

实验发现：TF–IDF 词袋模型结合随机森林算法能够更准确地表示文本特征，具有较好的表现，65.8% 的准确率已经接近可以接受的水准，可作为预测空管事件类型的首选方案。接下来，将对算法 2 的随机森林模型进行参数调优。

3.4.2　模型优化

通过网格搜索（Grid Search）的方法，调整随机森林的各项参数，在已有数据集上寻找最优参数组合。在随机森林模型中，最重要的参数有：森林的最大决策树个数 n_estimators、随机森林划分时考虑的最大特征数 max_features、决策树最大深度 max_depth、内部节点再划分所需最小样本数 min_samples_split、叶子节点最少样本数 min_samples_leaf 等，调优策略和结果如下。

n_estimators，分别取值从 50 到 100，步长为 10，实验最优取值为 60；

max_depth，分别取值从 3 到 15，步长为 2，实验最优取值为 13；

min_samples_split, 分别取值从 50 到 150，步长为 20，实验最优取值为 120；

min_samples_leaf, 分别取值从 10 到 150，步长为 20，实验最优取值为 90；

max_features, 分别取值从 3 到 15，步长为 2，实验最优取值为 11；

调优后准确率提升到 71%，相比调优之前提高了 5.2%，模型训练到这里已经难以提高，如需进一步提高模型的泛化能力，需要尝试输入更多空管事件的文本数据等方法。

3.5　自动化处理系统前端设计

通过 JQuery 和 Bootstrap 设计和开发安全信息机器处理的前端 HTML 页面，如图 2 所示，同时再修改 Python 脚本，并且在 c# 后端调用 Python 脚本，完成前

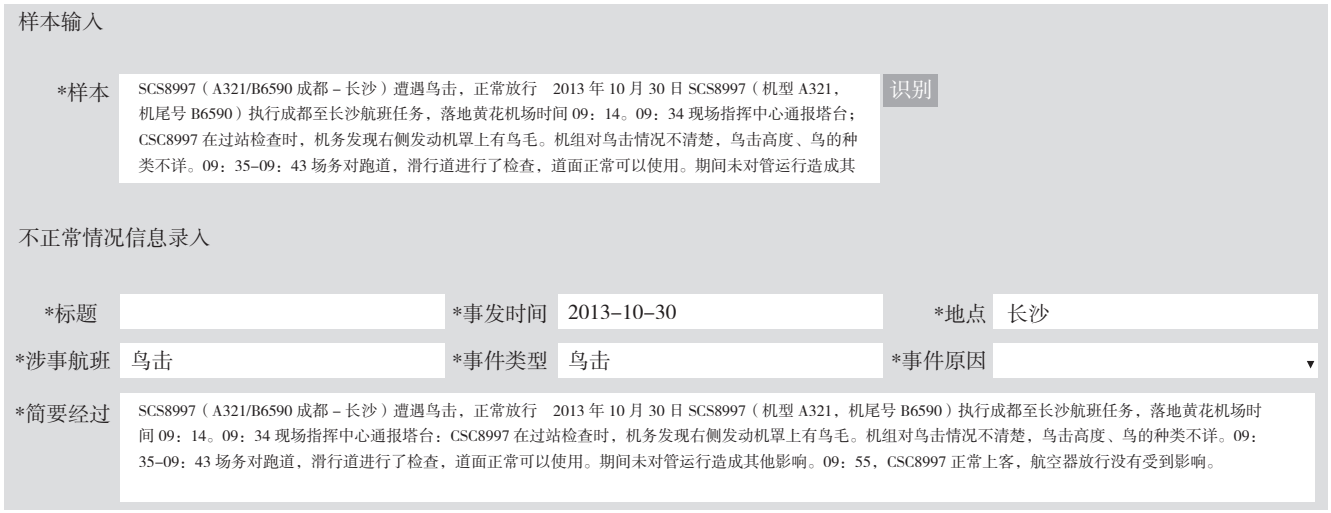

图 2　安全信息机器处理的用户页面

后端的整体开发。用户可以在机器学习识别文本一栏输入事件经过内容，点击识别按钮后，在相应栏目将自动显示分类结果，待用户确认无误后即可完成分类录入，大幅减少手工录入时间。

4 总结和展望

本文提出了一种自动处理空管安全信息的方法，创造了空管安全信息分词的词典，对行业安全信息管理有着积极的意义。通过词典分词和关键词映射关系的算法优化，有效提高了安全信息处理的准确率。空管行业可细分为如下专业：管制、情报、通信导航和航空气象，安全管理部门所掌握的管制专业安全信息数量最多，对应事件的处理准确率相对较高。非管制专业由于历史数据偏少，模型训练存在不足，应在后续工作中扩大非管制专业的数据搜集和模型训练，提高空管行业各专业安全信息处理的准确率。另外，如果尝试基于关键字的人工经验标注来提高模型的准确率，需要大量专家的经验提取和长时间的模型优化，其准确率的提升有待进一步研究。

空管安全信息分词词典的建设和应用填补了行业安全信息数据处理的空白，可广泛应用到空管运行品质监控、无线电通话语音识别等技术中，还可以扩展到自愿报告、综合安全信息的自动处理，结合无线电通话语音捕获和人工语音录入技术，提高安全管理部门对各类安全信息的处理能力，实现全天候、全要素的安全信息自动采集、识别和分析，助力“强安全”的四强空管建设。

参考文献

[1] 张聪俊 . 空管不安全事件报告机器分析关键方法研究 [D]. 天津：中国民航大学，2019.

[2] 田继存 . 文本分类及其在民航安全自愿报告分析中的应用研究 [D]. 天津 : 中国民航大学，2010.

[3] 崔振新，卢昊文 . 民航安全信息中实现关键词提取的方法 [J]. 交通信息与安全 .2016（05）.

[4] ZhangY,Jin R, Zhou Z H. Understanding bag-of-words model: a statistical framework[J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1-4): 43-52.

[5] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

[6] Ho T K. Random decision forests[C]//Proceedings of 3rd international conference on document analysis and recognition. IEEE, 1995, 1: 278-282.

（责任编辑　王立新）

（上接第 58 页）

1）广汉机场春季瞬时风速≥ 7 m/s 集中在午后；3 月瞬时风速≥ 7 m/s 频次出现最多为 16~17 时，4 月为 12~13 时，5 月则主要集中在 14 时前后，且 4~5 月夜间 19~20 时瞬时风速≥ 7 m/s 的频率较 3 月明显增加；

2）广汉机场雷暴次数日变化分布情况呈“U”型分布，雷暴大多出现在 06~11 时和 17~21 时段内；

3）广汉机场主导能见度≤ 1.6 km 和≤ 0.8 km 的次数逐月降低。在 3 月，主导能见度≤ 1.6 km 的现象在 06~13 时均有出现，且低能见度出现频次随时间先增加后减少；在 4 月，主导能见度≤ 1.6 km 的情况主要集中在 06~11 时，且低能见度出现次数随时间推移呈递减趋势；5 月仅出现一次，且持续时间较 3、4 月明显缩短。主导能见度≤ 0.8 km 情况在 3~4 月仅出现在 06~10 时，造成本场出场延迟；5 月仅出现在早间出场前，对按时出场无影响，且其余时段再无主导能见度≤ 0.8 km 情况出现。

4）广汉机场春季日降水次数自 3 月起逐月递增，近 5 年春季 06~22 时降水分布存在双峰值区：降水频次第一个峰值区在 3 月主要集中在 09 点，4、5 月则偏早，主要出现在 07 点，第二个大值区表现为 19~20 时。

参考文献

[1] 张燕光 . 航空气象学 [M]. 北京：中国民航出版社 , 2014.

（责任编辑　王立新）