doi:10.11835/j.issn.1000-582X.2025.216
引用格式:乔剑锋, 刘萱, 艾莉莎, 等. 基于SVM和归一化熵模型的隐患文本分类与类型特征分析[J]. 重庆大学学报, 2026,49(2): 105-115.

基于SVM和归一化熵模型的隐患文本分类与类型特征分析

乔剑锋[1],刘 萱[1],艾莉莎[2a,2b],张丽玮[1],王 汀[1]
(1. 首都经济贸易大学 管理工程学院,北京 100070; 2. 北京邮电大学 a.《北京邮电大学学报(自然版)》编辑部; b. 经济管理学院社会化网络信息研究中心,北京 100876)

摘要:为了提高隐患信息数据组织和检索的效率,支持更复杂的信息处理任务,需要采用有效技术手段对数据进行自动分类和类型分析。支持向量机(support vector machine,SVM)可以对自由文本进行自动分类,但是算法的工作原理是在训练集中寻找最优分类边界,不能发现类型典型特征。为了分析类型样本的共同特征,提出采用归一化熵模型寻找类型典型特征,改进当前词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)类型特征识别方法。以政府某应急管理局的2 534条执法检查记录为例,采用SVM进行自动分类,准确率高达97%。同时通过归一化熵模型给出各类型的典型特征,为制定隐患排查专项整治策略提供决策支持。实验结果表明,采用SVM和归一化熵模型的组合技术可以高效解决文本分类和类型特征识别的综合问题。

关键词:文本挖掘;数据挖掘;隐患排查;支持向量机;熵
中图分类号:X928 **文献标志码:**A **文章编号:**1000-582X(2026)02-105-11

Classifications and characterization of safety hazard texts

QIAO Jianfeng[1], *LIU Xuan*[1], *AI Lisha*[2a,2b], *ZHANG Liwei*[1], *WANG Ting*[1]
(1. School of Management Engineering, Capital University of Economics and Business, Beijing 100070, P. R. China; 2a. Editorial Department of Journal of Beijing University of Posts and Telecommunications (Nature Edition); 2b. Social Network Information Research Center, School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, P. R. China)

Abstract: To improve the efficiency of organizing and retrieving safety hazard information and to support more complex information processing tasks, effective technical methods for automatic text classification and type analysis are required. Support Vector Machine (SVM) can automatically classify unstructured text. However, their underlying principle focuses on identifying optimal classification boundaries within the training set and does not facilitate the extraction of representative features for each text category. To address this limitation, a normalized

收稿日期:2024-07-15 **网络出版日期:**2025-06-17
基金项目:中国高校科技期刊研究会专项基金项目(CUJS2024-GJ-A01)。
Supported by the Special Fund Project of the Society of China University Journals (CUJS2024-GJ-A01).
作者简介:乔剑锋(1977—),男,副教授,博士,主要从事安全数据挖掘以及安全风险预警和评价方向研究,(E-mail) qiaojianfeng@cueb.edu.cn。

entropy model is proposed to search for typical category features, thereby improving the traditional term frequency-inverse document frequency (TF-IDF) based feature recognition method. Using 2 534 law enforcement inspection records from a government emergency management bureau as a case study, SVM was used for automatic text classification and achieved an accuracy of up to 97%. Meanwhile, the normalized entropy model was used to extract representative features for each category, providing decision support for formulating targeted rectification strategies in hazard investigation. Experimental results show that the combined use of SVM and the normalized entropy model effectively addresses both text classification and category feature recognition tasks.

Keywords: text mining; data mining; hazard investigation; support vector machine; entropy

国家安全生产“十四五”规划指出强化事故隐患排查治理,深化工贸行业安全专项整治,强调安全隐患排查作为当前和今后一个时期我国安全生产工作的重中之重。要从源头上防范和化解重大安全风险,必须在事故发生之前彻底排查和治理各种潜在隐患[1]。随着工矿企业安全隐患排查工作的开展,隐患文本信息数据量不断累积,汇集成大数据,已经严重超出了分析人员的处理能力。如在面对类别复杂的隐患文本分类问题时,手动分类需要大量人工并对工作人员的知识背景也有较高要求。如何快速准确地对安全隐患文本进行自动分类,是当前安全叙述文本挖掘研究的热点问题。分类问题是认识客观事物的基础问题,避免因分类随意而模糊各类型规律的差异。当前安全隐患排查要求分类分级的精细化管理,研究隐患文本分类问题,对安全管理决策的科学建议具有重要意义。

文本分类作为自然语言处理的一个典型问题,被广泛应用于情感分析、新闻分类等方面,也被广泛应用于隐患文本、未遂事故文本和事故文本等3类安全叙事文本的分类中。Wellman等[2]最先将文本分类方法应用到安全叙事文本,随后,陈孝慈等[3]将文本分类技术应用到煤矿安全隐患文本,国内掀起安全叙事文本分类的研究热潮。结合安全叙事文本特点,当前主要开展扁平分类器(非层次分类)的研究,取得的研究成果包括:1)不同类型的安全叙事文本均开始进行自动分类研究,包括工人工伤赔偿报告[4]、医院急诊损伤事故[5]、民航空管隐患[6]等。研究案例数量从几百到十几万不等,如火灾事故案例有755个[7],安全生产氛围文本案例达12万[8]个。2)各类浅层学习和深度学习分类器已被广泛应用于安全叙事文本的自动分类研究中[9]。其中浅层学习方法包括支持向量机(support vector machine,SVM)[3]、线性回归(linear regression,LR)[4]、朴素贝叶斯(naive Bayes,NB)[7]和随机森林(random forest,RF)[10]。深度学习方法涵盖双向编码器表示(bidirectional encoder representations from transformers,Bert)[11]、卷积神经网络(convolutional neural network,CNN)和双向长短期记忆网络(bidirectional long short-term memory network,BiLSTM)[12]等。3)大部分自动分类性能指标大于0.9,说明自动分类方法可以应用于安全与应急管理领域的辅助决策。不同的分类标准和需要导致分类数量不同,少则2类[10],多则26类[4]。一般分类数量多会导致分类精度下降。

为了提高安全叙事文本自动分类性能,当前主要从4个方面开展研究。1)改善训练样本,提高训练样本包含的信息量。王洁宁等[6]采用过采样技术(synthetic minority oversampling technique,SMOTE),现有向量空间范围内,随机增加样本量。2)降低特征词空间维度,保留关键信息。尚麟宇等[13]采用卡方检验进行特征降维,也可以采用主要分分析(principal component analysis,PCA)技术。3)改进文本向量空间表示模型,提高关键特征词的权重系数。李华等[11]考虑不同隐患类别特征词应该具有不同的权重系数。4)选择合适的分类器并优化分类器的参数设置[9]。

采用深度学习分类器进行文本分类是研究热点,但是使用深度学习分类:1)需要进行词嵌入。即每一个词均需要用多达100维(或更多)向量进行表示,词向量训练需要大规模数据。2)需要构建深度学习分类器。它由多个模型级联构成而且每个模型都需要进行参数设置,为了找到最优参数,需要通过网格搜索(grid

search）确定最优参数。3）需要大量训练样本集。深度学习分类器需要大规模训练集，这样模型中的参数才能得到较为充分的学习训练。对于较少训练集，而且类型特征较为明显的样本集，浅层分类器仍然是最好的选择。浅层分类器占用计算机资源较少、训练时间短、分类精度高。Qiao等[9]以建筑事故为样本，对比浅层分类器和深度学习分类器的性能，发现支持向量机（support vector machine，SVM）（浅层分类器）分类性能最好，所以文中也采用SVM分类器。同时文中在实验部分，也测试了其他常用浅层分类器的分类性能。

识别出零散、重复、模糊的类型特征，是认知类型本质、掌握类型规律的关键环节。SVM的核心机制是在样本集中寻找最优分类边界，处于该边界上的样本实例即为支持向量。分类过程不涉及寻找类中心以及发现类型特征等环节。为了识别类型特征，研究学者采用文本表示方法TF（term frequency）[14]，TF-IDF（TF-inverse document frequency）[15-16]和卡方统计量[16-17]来识别类型特征。如果某类型中每个文档都包含某一特征词，而且特征词在不同文档中概率分布是一致的，通常可认为该特征词是类型的共同特征。在信息熵模型中，如果所有变量的概率分布均相同，则熵值达到最大值[18]。为识别不同类型的典型特征，文中提出基于归一化熵模型的TF-IDF类型特征识别方法，它考虑了特征词的类内分布信息，类型特征识别效率较高。

采用文本分类器进行安全叙事文本自动分类后，针对类型特征的研究相对较少，而且主要是采用TF-IDF进行类型特征识别，难以排除异常特征词的干扰。文中的核心贡献在于融合SVM与熵模型的各自优势，同步解决安全叙事文本的自动分类与类型特征识别问题。SVM虽具备优异的分类性能，但其核心原理是构建最优分类超平面，无法直接挖掘类型的典型特征。为此，文中在SVM分类结果的基础上，提出一种结合改进型归一化熵模型的TF-IDF特征识别方法，可有效识别文本类型的共同特征。

1　隐患文本

安全事故隐患记录人的不安全行为和物的不安全状态。下面简单介绍文中使用的隐患数据库、隐患类型分布和隐患文本特征，为隐患文本分类提供背景材料。

1.1　数据集

安全隐患文本或安全检查报告作为隐患语料集的数据来源，采用安全管理人员发现隐患并记录隐患的方式编写。数据集选取了某政府应急管理局2021年的2 534条执法检查记录，原始数据包括隐患记录日期、隐患描述、隐患单位、单位所属行业等信息，如表1所示，其中隐患描述就是文本分类中所指的文档对象。

表1　安全隐患原始数据信息

Table 1　Original data information of safety hazard

排查日期	隐患描述	隐患单位	行业
2021-06-11	未提供安全生产教育和培训的档案	X房地产经纪有限公司	房地产
2021-06-18	一层商户旁边安全出口疏散通道被分割为商户摊位	X商品批发市场有限公司	商业服务
2021-12-13	个别通道内堆放杂物	X科技有限公司	科技推广与应用服务业

1.2　手动分类统计

参照中华人民共和国应急管理部新颁布的《工贸企业重大事故隐患判定标准》，并结合现有隐患记录的实际情况，将安全隐患划分为2个一级类别和13个二级类别。一级标签包括基础管理类事故隐患及现场管理类事故隐患。其中基础管理类事故隐患包括资质证照、安全生产管理机构及人员、安全规章制度、安全培训教育、相关方管理、个体防护装备、职业健康、应急管理、隐患排查治理；现场管理类事故隐患包括作业场所、设备设施、原辅物料（产品）、安全技能等。对收集整理的隐患进行手动分类，结果如表2所示。

表2 安全隐患标签及样本分布

Table 2 Labels and sample distributions of the safety hazard text

类别	标签	数目	类别	标签	数目
基础管理类事故隐患	0资质证照	60	基础管理类事故隐患	7应急管理	210
	1安全生产管理机构及人员	22		8隐患排查治理	174
	2安全规章制度	298	现场管理类事故隐患	0作业场所	680
	3安全培训教育	291		1设备设施	529
	4相关方管理	16		2原辅物料、产品	72
	5个体防护装备	60		3安全技能	56
	6职业健康	66			

1.3 隐患文本特征

安全隐患文本具有4方面的特征:一是隐患文本属于自由描述文本,存在格式不统一、描述不规范、错字等问题。同时原始数据集还夹杂着数字、字母、符号、空格等不同格式信息,以及包含助词、副词、语气词等没有语义含义的词语。在文本挖掘前,需要进行文本清洗。二是隐患文本包含大量专业词汇。所以在中文分词过程中需要事先编辑好专业词汇表,通过有监督分词方法提高分词准确率和效率。三是隐患文本的文本长度相对较短。每条安全隐患的字符数一般少于30个,总体特征词空间维数也相对较小,导致关键特征词的区分度不足,增加了分类难度。四是隐患类别间数据分布不均衡。隐患出现频次的不同会导致各类别隐患数据量分布不均衡。在训练过程中小样本类型训练不充分,导致分类预测精度下降。

2 分类步骤和模型

文本分类通过带有类型标签的训练集(有监督学习),解决叙事文本自动分类问题。通常包括文本预处理、文本分词、文本表示、文本分类、性能评估和类型特征分析6个步骤(步骤6在第3节详述)。

2.1 文本预处理

文本预处理是将文本切割成具有实际含义的特征词(如字、词、词组),同时尽可能去除无用的信息(如停用词、标点符号等)。文本预处理主要包括文本清洗和去除停用词2个步骤。文本清洗是剔除文本中的标点符号、特殊字符、字母、数字等语料噪声或无用信息,同时不改变隐患描述的语义信息。去除停用词是指使用停用词表对原始文本进行清洗,同时去除文本中人名等与分类信息无关的特征词。

2.2 文本分词

在中文分词过程中,由于词是中文语义表达的基本单元,选取词作为特征要优于字和词组。文中采用结巴(Jieba)分词,并加载自定义安全隐患行业词表(从业人员、应急照明、职业病、资格证书、事故隐患等)实现中文自动分词。分词完成后总计有1 015特征(即文本分类的文本输入的特征词维度是1 015)。

2.3 文本表示

对于浅层分类器,经常使用文本向量空间表示模型(vector space model,VSM)对文本进行向量化,具有代表性的模型是TF-IDF。对于第i个特征词t_i在第j个文档d_j中的权重系数,为

$$T_{\text{TF-FIDF}}(t_i, d_j, D) = T_{\text{TF}}(t_i, d_j)\, I_{\text{IDF}}(t_i, D) \tag{1}$$

式中:$1 \leqslant i \leqslant m$,$m$是文本集$D$中包含的特征词数量,一般$m$值较大(如2.2节提及的1 015),降维处理就是删除一些不重要的特征词;$1 \leqslant j \leqslant n$,$n$是文本集$D$中包含文档的个数;$T_{\text{TF}}(t_i, d_j)$表示$t_i$在$d_j$中的数量,出现的次数越多该词就越重要。

$$I_{\text{IDF}}(t_i, D) = \log \frac{1+n}{1+D(t_i)} + 1\,, \tag{2}$$

式中：$D(t_i)$表示整体文档集D中包含t_i的文档数目。如果大部分文档均出现t_i，说明t_i不是文档的特有特征，权重系数要相对降低一些。式(2)最后部分加1是使I_{IDF}在极值情况下不为0。每个文档向量化后是一个1行m列的向量，其中每列元素的数值就是$T_{\text{TF-FIDF}}(t_i,d_j,D)$，它表达$t_i$特征词对$d_j$文档的重要程度。

2.4　SVM分类器

SVM是由Vapnik领导的AT&T Bell实验室研究小组在1995年提出的一种基于统计学习理论的分类方法。遵循结构风险最小化原则，不仅考虑最小化错误分类的数量，还考虑模型的复杂度，同时具有较好的泛化能力，在未见过的数据上的表现通常比那些只关注最小化训练误差的模型要好[19]。在处理小样本、非线性和高维度模式识别分类问题时，具备独特的优势[20]。其工作原理是假设样本线性可分，对于二分类问题，就是寻找最优超平面，该平面不但能将2类样本点正确分开，而且使最靠近超平面的2类样本点到超平面的距离之和最大。假设最优超平面H^0，平面法向量$\boldsymbol{w}$，偏移量b。H^1是类型1的边界平面，H^{-1}是类型2的边界平面，两平面都平行于H^0，在H^1和H^{-1}上的样本数据称为支持向量。2类样本数据之间的最大值就是$2/\|\boldsymbol{w}^2\|$。分类问题是求解如下最优化问题，见式(3)和式(4)。

$$\min_{w,b}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}, \tag{3}$$

$$\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\times x_i+b\right)-1\geqslant 0,\ 1\leqslant i\leqslant n。 \tag{4}$$

式中：x_i是输入特征向量；y_i是对应的类别标签。假设有n对训练集(x_i,y_i)，并且$w,x_i\in R^d$，$y_i=\pm 1$。采用拉格朗日乘子法进行对偶优化，可求得其最优解[21]，得出的分类函数为：

$$f(x)=\operatorname{sgn}(\boldsymbol{w}^{\mathrm{T}^*}\times x+b^*), \tag{5}$$

$$w^*=\sum_{i,j=1}^{m}y_i\alpha_i^*x_i。 \tag{6}$$

式中，$\alpha_i^*\geqslant 0$是拉格朗日系数，大部分都是等于0，只有支持向量的拉格朗日系数大于零。体现文本特征表示稀疏性问题和SVM需要样本量少的特点。

在面对多分类问题时(大于2个分类)，采用拆解法求解，即把一个多分类任务分成多个二分类任务解决。对拆分出来的每一个二分类任务进行训练。最后在进行预测的时候，把这些二分类学习器预测结果进行集成来获得多分类问题结果。

2.5　分类性能指标

分类问题一般是多分类问题，即类型数目大于2。分类预测结果一般采用混淆矩阵(confusion matrix)来表示，并基于混淆矩阵计算分类性能指标。混淆矩阵实质上是一个$M\times M$的矩阵(M是类型数)，用$\boldsymbol{C}'$表示，其中矩阵元素是$c'_{ij}(i,j=1,2,\cdots,M)$。$c'_{ij}$表示属于第$i$个类的样本被预测到第$j$类的样本数。假定样本数为$N$，则有$N=\sum_{i,j=1}^{M}c'_{ij}$。最理想的预测是

$$\begin{cases}c'_{ij}=c_i, & i\neq j;\\ c'_{ij}=0, & i=j。\end{cases}$$

式中，c_i是第i类(实际标签)的样本总数。

对于某类型的分类性能指标，经常使用3类性能指标来度量，精确率p_i(precision)，召回率r_i(recall，或灵敏度sensitivity)和F1值f_i(F1 score)。计算公式分别如下：

$$p_i=\frac{c'_{ii}}{\sum_{j=1}^{M}c'_{ji}}=\frac{c'_{ij}}{c'_{\cdot i}}=\frac{\boldsymbol{tp}_i}{\boldsymbol{tp}_i+\boldsymbol{fp}_i}, \tag{7}$$

$$r_i=\frac{c'_{ii}}{\sum_{j=1}^{M}c'_{ij}}=\frac{c'_{ii}}{c'_{i\cdot}}=\frac{\boldsymbol{tp}_i}{\boldsymbol{tp}_i+\boldsymbol{fn}_i}, \tag{8}$$

$$f_i=\frac{2p_ir_i}{p_i+r_i}=\frac{\boldsymbol{tp}_i}{\boldsymbol{tp}_i+\frac{1}{2}(\boldsymbol{fp}_i+\boldsymbol{fn}_i)}。 \tag{9}$$

从混淆矩阵的列向量考虑(即预测样本标签),有$c'_{.i}$个样本被预测为i个类,其中有$\boldsymbol{tp}_i=c'_{ii}$个样本实际标签也是第$i$个类(正确预测),有$\boldsymbol{fp}_i=c'_{.i}-c'_{ii}$个样本被预测为其他类(错误预测)。所以第$i$类的精确率指样本本身属于第$i$类并且又被预测为第$i$类的样本数与被预测为第$i$类的样本总数的比率。

从混淆矩阵的行向量考虑(即实际样本标签),有$c'_{i.}$个样本属于第i个类($c'_{i.}=c_i$),其中有$tp_i=c'_{ii}$个样本也预测到了第i个类(正确预测),有$\boldsymbol{fn}_i=c'_{.i}-c'_{ii}$个样本被预测为其他类(错误预测)。第$i$类的召回率指属于第$i$类的样本并且被预测为第$i$类与属于第$i$类样本的比率。

F1值是精确率和召回率的调和平均。每个类都有p_i、r_i、f_i,这些指标会汇总到分类报告中。对于整体分类性能指标,一般采用准确度F_{accuracy}(accuracy)和平均F1值$F_{\text{f1-score}}$(average F1 score)来度量。准确度表示被正确分类的样本数(所有类型)占总样本数。平均F1值是所有类型的F1值的平均值。平均F1值一般不适合评价类型样本数分布不均衡的情况,因为含样本数少的类型的F1值较低(虽然只占少数样本),整体上会降低平均F1值。因此,可引入加权灵敏度$W_{\text{avg-sen}}$(weighted average sensitivity)作为替代指标,如式(12),其数学实质与准确度等价。一般不使用加权F1值,因为f_i是调和平均值,加权后再次调和,公式复杂并且不容易被直观理解[9]。

$$F_{\text{accuracy}}=\frac{\sum_{i=1}^{M}c'_{ii}}{N}, \tag{10}$$

$$F_{\text{f1-score}}=\frac{\sum_{i=1}^{M}f_i}{M}, \tag{11}$$

$$W_{\text{avg-sen}}=\frac{\sum_{i=1}^{M}.r_i c'_{i.}}{N}。 \tag{12}$$

3 类型特征分析方法

在进行自动分类时,类型特征分析是核心步骤之一。在自动分类器训练之前,通过类型特征分析进行特征选取,可以提高分类性能;在分类完成后,也可通过类型特征分析,抽取关键特征来认识类型和区分类型。在自动分类工作完成的基础上,结合类型特征分析,可精准识别各类型的典型特征。

3.1 传统类型特征分析方法

类型特征可以通过类中心权重系数较高的特征词来表征。由于文档集D中的每个词均可以用$T_{\text{TF-FIDF}}(t_i, d_j, D)$表示,所以某一类型的特征词可以用该类型所包含的所有文档D_c的平均$T_{\text{TF-FIDF}}(t_i, d_j, D)$值较高的特征词来表示,见式(13)。这也是基于类中心分类器的设计原理。但是这类采用平均$T_{\text{TF-FIDF}}(t_i, d_j, D)$来分析类中心的方法难以处理数据集中的异常点,如$t_i$仅在$d_j$中出现,而且$T_{\text{TF-FIDF}}(t_i, d_j, D)$值较高,同时$D_c$样本少,有可能导致平均$T_{\text{TF-FIDF}}(t_i, d_j, D)$较高,但$t_i$并不是类型典型特征。所以,一般基于类中心设计的分类器其分类性能不是很理想。

$$\overline{T_{\text{TF-FIDF}}(t_i, d_j, D_c)}=\frac{\sum_{i=1}^{n}T_{\text{TF-FIDF}}(t_i, d_j, D)}{n}。 \tag{13}$$

3.2 基于归一化熵模型的特征分析方法

文中引入一种新的类型特征词权重系数表示方法,即信息熵模型。信息熵为一个物理概念,最早是由克劳修斯在热力学中提出的,用以描述系统的状态。而后信息熵被引入多个领域,从而产生了玻尔兹曼熵、信息熵、概率测度熵等[22]。假设某类型包含n个文档,针对某一特征词,该特征词在某一类型中的信息熵为

$$H=-\sum_{i=1}^{n}p_i\log p_i, \tag{14}$$

式中:p_i表示特征词在第i个文档的$T_{\text{TF-FIDF}}$值与类型样本集中的总$T_{\text{TF-FIDF}}$的比率,即

$$p_i = T_{\mathrm{TF}}(t_i, d_j, D) / \sum_{i=1}^{n} T_{\mathrm{TF}}(t_i, d_j, D)。\tag{15}$$

使用熵模型表示特征词权重系数有 3 个优点：1）包含信息量大。这也是 $T_{\text{TF-FIDF}}(t_i, d_j, D)$ 方法的优点，包含文档本身信息的信息（term frequency，TF），同时也包含总体样本的信息（inverse document frequency，IDF）。2）能反映类型的共同特征。如果类型中的每个文档都包含同一特征词，而且特征词在不同文档中的权重系数是一致的，说明该特征词是类型的共同特征。在熵模型中，如果所有 p_i 都相同，则熵达到最大值。3）有效去除类型中稀有特征词的干扰。如果类型中所有文档只有一个文档包含某一特征词，如果采用平均 $T_{\text{TF-FIDF}}(t_i, d_j, D)$ 计算，则该特征词在类型中的权重系数不等于 0；而采用熵模型计算，熵值为 0，即采用熵模型计算，稀有（或异常）特征词在类型中的权重系数会降低甚至为 0。

分类问题中不同类型样本数量往往不相等，部分类型样本数较少。而熵模型随着样本数的增加熵值增大，为了解决不同类型之间的特征词在同一水平下的比较问题，采用归一化熵来衡量类型特征的权重系数，见式（16），所有类型特征词的权重系数均为 $[0,1]$[23]。

$$H = -\sum_{i=1}^{n} p_i \lg p_i / \lg(n)。\tag{16}$$

4　实验结果

实验采用 Python 语言开发，使用了自然语言开发工具包 nltk 库和 scikit 库，对获取的隐患文本，按步骤进行自动分类和类型特征分析。

4.1　词云图

借助词云图可以直观地为管理人员展示隐患文本中的主要隐患。文本清洗和 Jieba 分词后，选取词频数较多的前 100 个关键特征词绘制隐患词云图，如图 1 所示。图中字体越大的特征词代表该词出现的频率越高。安全隐患文本中出现频率前 20 的特征词为：“记录”“制度”“出口”“培训”“应急”“杂物”“堆放”“教育”“设置”“疏散”“事故隐患”“改正”“逾期”“缺少”“标识”“排查”“通道”“演练”“指示”“救援”“从业人员”。这些特征词可以帮助安全管理者更好地理解安全隐患文本的关键主题和内容。

图 1　隐患词云图

Fig. 1　Cloud map of safety hazard words

4.2　SVM 分类结果

安全隐患数据按照训练集占 75% 和测试集占 25% 的比例进行随机抽取。经过 SVM 训练后，测试集预测结果见表 3 所示，样本整体分类的准确度为 97%。其中训练样本数少的类型的预测精度相对较差，如类型“1”。由于训练样本少，分类器无法学习到足够多的特征，从而导致无法对新数据进行准确的预测分类。

表3 各类别的分类准确度

Table 3 Accuracy of different classes

类型	Precision	Recall	F1	样本数	类型	Precision	Recall	F1	样本数
0	0.94	1.00	0.97	15	7	0.98	0.98	0.98	53
1	0.83	1.00	0.91	5	8	1.00	1.00	1.00	54
2	0.98	0.97	0.98	65	9	0.97	0.98	0.97	176
3	0.99	1.00	0.99	67	10	0.95	0.95	0.95	126
4	1.00	0.86	0.92	7	11	1.00	1.00	1.00	19
5	1.00	0.67	0.80	17	12	1.00	1.00	1.00	18
6	1.00	0.95	0.97	20	Accuracy	–	–	0.97	634

注:样本数表示测试集样本数。

SVM作为一种常用的分类算法,在文本分类任务中具有以下优势:1)SVM在文本分类任务中表现出较高的精度,特别是在小样本情况下表现更为突出。2)可以处理高维特征,在文本分类任务中,一篇文本往往被表示成高维空间向量,且向量具有稀疏性。SVM技术能在这个高维空间内区分不同类别的文本。3)适用于多分类任务,SVM不仅适用于二分类问题,也可以扩展到多分类问题。4)抗噪能力强,由于SVM采用边界学习策略,不受噪声干扰数据的影响。

4.3 类型特征

采用归一化熵模型分析各类的典型特征,表4中熵模型抽取的特征词反映各类型特征的熵值最大的前10个特征词,这些特征词反映了各类型的典型特征。

例如,“1.1资质证照”的特征词反映企业特种作业人员未持证上岗等典型隐患案例。再比如,企业“作业场所”和“设备设施”存在安全隐患较多(见表2),结合类型典型特征词表(见表4),发现作业场所主要存在安全出口堆放杂物,疏散指示标识设置等问题。设备设施主要存在临时用电、防水插座、应急照明灯等问题。这些安全隐患为将来制定隐患排查专项整治的策略提供决策支持。另外,从隐患所属行业的角度进行分析,发现餐饮场所隐患较多,行业安全管理水平较低。典型隐患包括后厨未安装防水插座、疏散通道堆放杂物等。即餐饮企业需要加强隐患排查治理,同时政府部门需要加大检查频次和处罚力度,并加强安全宣传教育。

表4 类型特征抽取对比

Table 4 Comparison of extracted feature words

编号	类型名称	平均TF-IDF抽取的特征词	熵模型抽取的特征词	熵模型抽取的特有特征词代表的典型隐患举例说明
0	1.1资质证照	作业、电工、特种、人员、证件、原件、未见、上岗、持证、出示	作业、人员、特种、电工上岗、证件、未见、原件、出示、资格证书	特种作业人员未取得特种作业操作资格证书
1	1.2安全生产管理机构及人员	值班、高压、配电室、空调、临街、警戒线、单人、无人、化学品、护栏	值班、配电室、高压、空调、设置、专人、专用、临街、仓库、保管	(危险化学品)专用仓库未设专人负责管理
2	1.3安全规章制度	健全、生产、安全、制度、责任、责任制、记录、操作规程、机械设备、生产例会	安全、生产、健全、制度、记录、责任、责任制、操作规程、机械设备、生产例会	无(两模型抽取特征词一致)
3	1.4安全培训教育	培训、教育、生产、安全、从业人员、记录、档案、上岗、制度、情况	培训、安全、教育、生产、从业人员、记录、档案、健全、制度、情况	未健全安全生产教育培训制度及培训档案

续表 4

编号	类型名称	平均TF-IDF抽取的特征词	熵模型抽取的特征词	熵模型抽取的特有特征词代表的典型隐患举例说明
4	1.5相关方管理	协议、签订、管理、承包、单位、承包方、盐府、安全、交底、技术、合同	安全、协议、管理、签订、承包单位、承包方、盐府、合同、生产、改正	逾期未改正
5	1.6个体防护装备	防护用品、劳动、配电室、提供、国家标准、行业标准、符合、未回、送检、绝缘	防护用品、劳动、符合、配电室、提供、从业人员、国家标准、行业标准、改正、逾期	未为从业人员提供符合国家标准或者行业标准的劳动防护用品,或逾期未改正
6	1.7职业健康	危害、职业病、用人单位、职业、工作、场所、检测、申报、因素、进行	职业病、危害、用人单位、工作、场所、申报、规定、进行、职业、安全	用人单位未按照规定……
7	1.8应急管理	应急、演练、救援、预案、生产、记录、安全、健全、安全事故、出示	应急、演练、生产、救援、安全、记录、预案、健全、安全事故、制定	未按照规定制定生产安全事故应急救援预案
8	1.9隐患排查治理	事故隐患、排查、治理、制度、生产、健全、巡查、安全、记录、建立	事故隐患、排查、制度、安全、生产、治理、健全、记录、巡查、建立	无(两模型抽取特征词一致)
9	2.1作业场所	杂物、堆放、出口、疏散、安全、通道、指示、标识、设置、标志	安全、杂物、堆放、疏散、出口、设置、标识、指示、通道、标志	无(两模型抽取特征词一致)
10	2.2设备设施	防水、临时、照明灯、用电、插座、线路、保护装置、符合规范、防爆灯、要求	临时、防水、线路、要求、符合规范、用电、应急、插座、照明灯、设置	应急照明灯破损,未设置防爆灯
11	2.3原辅物料产品	码放、货物、化学品、过高、符合规范、仓库、说明书、混乱、物品、危险	码放、货物、化学品、符合规范、过高、使用、仓库、要求、一层、地下	地下一层违反规定储存和使用化学品
12	2.4安全技能	开启、方向、出口、正确、安全、固定、上锁、室内、配电、游泳馆	安全、出口、开启、方向、正确、配电、室内、固定、使用、上锁	安全出口使用电动推拉门

注:特征词按照权重系数从大到小排序,斜体字表示两模型的差异特征词,加粗特征词表示来自熵模型。

4.4　对比讨论

针对自动分类问题,研究学者主要关注2方面的问题:1)通过SMOTE和PCA技术看是否可以提高分类性能。通过实验,2种方法均没有提高隐患文本的分类性能。首先,SMOTE是对含有训练样本数少的类型通过随机插值产生新的样本;算法本身不能在训练样本中添加新的特征词,只能在类型原有特征词的基础上随机改变特征词的权重系数:加入SMOTE并没有带来显著的改进。其次,采用PCA将高维数据映射到低维空间,可以提高模型的运行效率并减少过拟合的风险,但是由于维数降低,导致部分特征缺失,从而使模型性能下降。2)文中把经常使用的10个浅层学习分类器的性能进行对比(采用scikit-library中函数实现),见表5所示,其中SVM的分类性能最好;GB、RF、SNN等分类器性能也较好,准确率达到96%,主要由于样本噪音数据少并且线性可分;基于类中心原理的NC分类器性能较低,准确率为87%,其原因在3.1节中进行了讨论。

对比熵模型和TF-IDF关于类型特征的抽取效率,计算类型各特征词的权重系数,进行特征抽取,按照权重系数从高到低选择排名前10的特征词代表类型特征。如表4所示,把基于TF-IDF和归一化熵模型抽取的特征词进行对比。通过两模型的对比也相互验证两模型的有效性,同时发现熵模型抽取效率更高。表中把基于熵模型抽取的独特特征词,给出其对应的典型隐患,说明其特征抽取的有效性。

表5 浅层分类器性能比较

Table 5 Performance comparison of shallow classifiers

分类器名	分类器函数	准确率	分类器名	分类器函数	准确率
SVM	LinearSVC	0.97	NB	MultinomialNB	0.88
NC	NearestCentroid	0.87	DT	DecisionTreeClassifier	0.95
KNC	KNeighborsClassifier	0.93	RF	RandomForestClassifier	0.96
GB	GradientBoostingClassifier	0.96	LR	LogisticRegression	0.95
Bagging	BaggingClassifier	0.95	SNN	MLPClassifier	0.96

注：NC，nearest centroid；KNC，K-neighbors classifier；GB，gradient boosting；DT，decision tree；SNN，shallow neural network。

5 结 论

隐患排查结果主要表现形式是隐患文本，为有效遏制安全事故，需要建立数据驱动、以自然语言理解为核心的认知计算模型，形成从大数据到知识、从知识到决策的能力。自动分类技术可以实现隐患类型的自动统计分析，并且结合类型特征分析可以为深化工贸行业安全隐患排查专项整治提供决策支持。

1）以政府隐患排查文本为例，使用Python自动分类软件包验证了线性SVM的分类性能表现最好，准确度达到了0.97，隐患类型可以用类型典型特征表示（线形可分）。

2）采用归一化熵模型分析类型中各特征词的权重系数，抽取权重系数较高的特征词分析类型的共同特征，为进一步认识类型和抽取类型特征提供依据。

3）采用SVM和归一化熵模型的组合技术可以高效解决文本分类和类型特征识别的综合问题。以安全隐患文本分类与类型特征分析为例，验证组合技术的有效性。

参考文献

[1] 宋守信, 陈明利, 翟怀远, 等. 新修《安全生产法》中的安全发展理念: 从条款第三条谈起[J]. 安全, 2021, 42(11): 10-14, 9.
Song S X, Chen M L, Zhai H Y, et al. Safety evelopment concept in the work safety law of 2021 amendment version: talking from article 3 of the work safety law[J]. Safety & Security, 2021, 42(11): 10-14, 9. (in Chinese)

[2] Wellman H M, Lehto M R, Sorock G S, et al. Computerized coding of injury narrative data from the National Health Interview Survey[J]. Accident; Analysis and Prevention, 2004, 36(2): 165-171.

[3] 陈孝慈, 谭章禄, 单斐, 等. 基于Bigram的安全隐患文本分类研究[J]. 中国安全科学学报, 2017, 27(8): 156-161.
Chen X C, Tan Z L, Shan F, et al. Research on text categorization for hidden dangers based on Bigram[J]. China Safety Science Journal, 2017, 27(8): 156-161. (in Chinese)

[4] Marucci-Wellman H R, Corns H L, Lehto M R. Classifying injury narratives of large administrative databases for surveillance: a practical approach combining machine learning ensembles and human review[J]. Accident Analysis & Prevention, 2017, 98: 359-371.

[5] Nanda G, Vallmuur K, Lehto M. Intelligent human-machine approaches for assigning groups of injury codes to accident narratives[J]. Safety Science, 2020, 125: 104585.

[6] 王洁宁, 侯海洋, 贾奇. 不均衡空管危险源自由文本分类模型[J]. 安全与环境学报, 2022, 22(2): 826-835.
Wang J N, Hou H Y, Jia Q. Free text classification model for unbalanced air traffic management hazard reports[J]. Journal of Safety and Environment, 2022, 22(2): 826-835. (in Chinese)

[7] 葛继科, 陈栋, 王文和, 等. 基于改进朴素贝叶斯分类算法的火灾分类[J]. 安全与环境学报, 2019, 19(4): 1122-1127.
Ge J K, Chen D, Wang W H, et al. Fire classification based on improved naive Bayesian classification algorithm[J]. Journal of Safety and Environment, 2019, 19(4): 1122-1127. (in Chinese)

[8] 谢汉青, 邱少辉, 王寓霖, 等. 面向非均衡文本信息的企业生产安全氛围智能感知模型[J]. 安全与环境工程, 2022, 29(3): 47-54.
Xie H Q, Qiu S H, Wang Y L, et al. Intelligent perception model of enterprise production safety climate oriented to unbalanced

text[J]. Safety and Environmental Engineering, 2022, 29(3): 47-54. (in Chinese)

[9] Qiao J F, Wang C F, Guan S, et al. Construction-accident narrative classification using shallow and deep learning[J]. Journal of Construction Engineering and Management, 2022, 148(9): 04022088.

[10] Zermane A, Mohd Tohir M Z, Zermane H, et al. Predicting fatal fall from heights accidents using random forest classification machine learning model[J]. Safety Science, 2023, 159: 106023.

[11] 李华, 陈俞源, 高红, 等 . 基于改进Bert模型的建筑事故隐患分类方法研究[J]. 安全与环境学报, 2022, 22(3): 1421-1429.
Li H, Chen Y Y, Gao H, et al. Research on hidden danger classification method of construction accident based on improved Bert model[J]. Journal of Safety and Environment, 2022, 22(3): 1421-1429. (in Chinese)

[12] 刘斐, 文中, 吴艺 . 基于 BERT-BILSTM-CRF 模型的电力行业事故文本智能分析[J]. 中国安全生产科学技术, 2023, 19(1): 209-215.
Liu F, Wen Z, Wu Y. Intelligent analysis on text of power industry accident based on BERT-BILSTM-CRF model[J]. Journal of Safety Science and Technology, 2023, 19(1): 209-215. (in Chinese)

[13] 尚麟宇, 尹明, 肖畅, 等 . 基于BLS的铁路安全事件文本分类研究[J]. 中国安全科学学报, 2022, 32(6): 103-108.
Shang L Y, Yin M, Xiao C, et al. Research on text classification of railway safety incidents based on BLS[J]. China Safety Science Journal, 2022, 32(6): 103-108. (in Chinese)

[14] 张伟, 石倩, 何霄, 等 . 改进的TF-IDF算法在文本分类中的研究[J]. 信息技术与网络安全, 2021, 40(7): 72-76, 83.
Zhang W, Shi Q, He X, et al. Research on improved TF-IDF algorithm in text classification[J]. Information Techology and Network Security, 2021, 40(7): 72-76, 83. (in Chinese)

[15] 田水承, 王雪晨, 范彬彬 . 基于文本挖掘的建筑施工坍塌事故致因研究[J]. 西安科技大学学报, 2022, 42(5): 849-855.
Tian S C, Wang X C, Fan B B. Research on causes of collapse accidents in building construction based on text mining[J]. Journal of Xi'an University of Science and Technology, 2022, 42(5): 849-855. (in Chinese)

[16] 陈志远, 王铁骊 . 基于文本挖掘和复杂网络的事故致因重要度评估方法: 以房屋市政较大以上事故为例[J]. 中国安全生产科学技术, 2022, 18(4): 224-230.
Chen Z Y, Wang T L. Evaluation method of accident causes importance based on text mining and complex network: a case study of larger and above accidents in housing and municipal engineering[J]. Journal of Safety Science and Technology, 2022, 18(4): 224-230. (in Chinese)

[17] 李珏, 王幼芳 . 基于文本挖掘的建筑施工高处坠落事故致因网络分析[J]. 安全与环境学报, 2020, 20(4): 1284-1290.
Li J, Wang Y F. Causation network analysis of the construction falling or collapsing accidents based on the text mining[J]. Journal of Safety and Environment, 2020, 20(4): 1284-1290. (in Chinese)

[18] 刘天雄, 陈辉华, 李瑚均 . 复合地层盾构施工安全影响因素及安全事故致因机理[J]. 铁道科学与工程学报, 2020, 17(1): 266-272.
Liu T X, Chen H H, Li H J. Research on safety impact factors and safety accident causation mechanism of subway shield construction in mix-ground[J]. Journal of Railway Science and Engineering, 2020, 17(1): 266-272. (in Chinese)

[19] 马创, 王尧, 李林峰 . 基于遗传算法与支持向量机的水质预测模型[J]. 重庆大学学报, 2021, 44(7): 108-114.
Ma C, Wang Y, Li L F. A water quality prediction model based on genetic algorithm and SVM[J]. Journal of Chongqing University, 2021, 44(7): 108-114. (in Chinese)

[20] 赵江平, 王垚 . 基于图像识别技术的不安全行为识别[J]. 安全与环境工程, 2020, 27(1): 158-165.
Zhao J P, Wang Y. Unsafe behavior recognition based on image recognition technology[J]. Safety and Environmental Engineering, 2020, 27(1): 158-165. (in Chinese)

[21] 张莹, 郭红梅, 尹文刚, 等 . 基于特征提取的SVM图像分类技术的无人机遥感建筑物震害识别应用研究[J]. 灾害学, 2022, 37(4): 30-36, 56.
Zhang Y, Guo H M, Yin W G, et al. Application of SVM image classification technology based on feature extraction in seismic damage identification of buildings by UAV remote sensing[J]. Journal of Catastrophology, 2022, 37(4): 30-36, 56. (in Chinese)

[22] 李彦苍, 王旭 . 基于信息熵的改进海豚群算法及其桁架优化[J]. 重庆大学学报, 2019, 42(5): 76-85.
Li Y C, Wang X. Improved dolphin swarm algorithm based on information entropy and its truss optimization[J]. Journal of Chongqing University, 2019, 42(5): 76-85. (in Chinese)

[23] Qiao J F, Li Y. Resource leveling using normalized entropy and relative entropy[J]. Automation in Construction, 2018, 87: 263-272.

（编辑　詹燕平）