郭心全, 吴霞, 李俊波, 等. 基于 GBDT 模型的接触网异物分类研究[J]. 智能计算机与应用,2024,14(6):41-49. DOI:10.20169/j.issn.2095-2163.240606

# 基于 GBDT 模型的接触网异物分类研究

郭心全, 吴 霞, 李俊波, 沈 鹃, 郝贵才
(中国铁道科学研究院集团有限公司 电子计算技术研究所, 北京 100081)

**摘 要**: 为解决铁路接触网异物信息文本数据利用不充分的问题,快速高效地判识接触网异物类别,开展接触网异物分类研究。首先,通过分析接触网异物文本特点,抽取出与接触网异物类别相关的实体建立“接触网异物词典”;其次,以 Jieba 分词工具加载该词典对文本数据进行分词并清洗;随后,通过词频-逆向文件频率(TF-IDF)算法挖掘文本信息的关键特征,并以 8:2 比例拆分训练集和测试集;最后,构建梯度提升决策树(GBDT)分类模型以训练集进行训练,以训练好的模型和测试集进行模型验证,并通过实验对比常用的 K 最近邻(KNN)、多项式朴素贝叶斯(MNB)、逻辑回归(LR)、随机森林(RF)、决策树(DT)等 7 个多类别文本分类模型。实验结果表明,基于 TF-IDF+GBDT 的接触网异物分类模型的精确率、召回率和 *F*1 值分别达到了 94.70%、94.74%和 94.53%,优于相比较的其他分类模型,具备一定的推广和应用价值。

**关键词**: 接触网; 异物; 文本分类; TF-IDF 算法; GBDT 模型
**中图分类号**: U226.8 **文献标志码**: A **文章编号**: 2095-2163(2024)06-0041-09

# Research on intrusion foreign objects classification of contact networks based on GBDT model

GUO Xinquan, WU Xia, LI Junbo, SHEN Kun, HAO Guicai
(Institute of Computing Technologies, China Academy of Railway Sciences Co., Ltd., Beijing 100081, China)

**Abstract**: To solve the problem of insufficient utilization of text data for foreign object information in railway contact networks, and to quickly and efficiently identify the types of foreign objects in contact networks, research on foreign object classification in contact networks has been carried out. Firstly, by analyzing the text characteristics of foreign objects in the contact network, entities related to the category of foreign objects in the contact network are extracted to establish a " dictionary of foreign objects in the contact network"; Secondly, use the Jieba word segmentation tool to load the dictionary to segment and clean the text data; Subsequently, the key features of text information are extracted using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, and the training and testing sets are split in an 8:2 ratio; Finally, a Gradient Boosting Decision Tree (GBDT) classification model is constructed to train the training set, and the trained model and test set are used for model validation. Seven commonly used multi class text classification models, namely K-Nearest Neighbor (KNN), Multinomial Naive Bayes (MNB), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), etc, are compared through experiments. The experimental results show that the accuracy, recall, and *F*1 values of the contact network foreign object classification model based on TF-IDF+GBDT reach 94.70%, 94.74%, and 94.53%, respectively, which is superior to other classification models compared to other models and has certain promotion and application value.

**Key words**: contact network; foreign objects; text classification; TF-IDF algorithm; GBDT model

## 0 引 言

电气化铁路是指以电力机车作为牵引动力的铁路,有着行驶速度快、运载能力强、无污染等特点。随着国内经济的快速发展,电气化铁路发展迅猛,截止 2020 年,国内电气化铁路运营里程约占铁路总里程的 74%[1-2]。接触网作为电气化铁路的重要组成部分,长期暴露于空气中,容易受到一些异物侵袭,如编织袋、风筝、塑料袋等,从而影响铁路行车安全[3]。而不同的异物对接触网的影响程度不同,但

**基金项目**: 中国铁道科学研究院集团有限公司科研项目(2022YJ067)。
**作者简介**: 郭心全(1994-),男,硕士,助理研究员,主要研究方向:铁路供电,工务信息化智能化。Email: gxq_2019@163.com
**收稿日期**: 2023-12-08

是都需要快速判别异物类别并进行妥善处理,保证电气化铁路供电畅通,确保铁路运输安全。因此,如何快速准确地判识接触网异物至关重要。

目前,大数据和人工智能技术已经逐渐渗透到各行各业,在铁路行业也得到了广泛应用,人工智能技术与铁路专业相结合的概念相继被提出[4-5],如智能铁路、智慧车站、智慧基础设施等。在人工智能技术中,图像处理技术相对于自然语言处理技术较为成熟,在各行业落地应用较多。很多铁路方面的研究课题大多是将图像处理技术应用于铁路车务、客运、工务、供电等专业,基于图像处理的接触网异物识别研究取得了不少成果,如徐伟等学者[6]基于接触网安全巡检装置(2C)采集图像数据,提出了高铁接触网异物自动化智能检测方法,以实现稳健、可靠、精准的高铁接触网安全异常检测。王科理等学者[7]提出一种基于深度学习的鸟巢异物检测方法,能够进行铁路鸟巢异物的有效检测,大大降低了人工干预的成本。Tan 等学者[8]针对绝缘子破损、污垢、异物和闪络四类主要缺陷提出了基于 Mask R-CNN 和多特征聚类模型的识别方法,并取得了较高的缺陷识别精确度。而在铁路文本数据方面也开展了一定的探索研究,如杨连报等学者[9]针对铁路信号设备不平衡故障文本数据,提出了基于文本挖掘的铁路信号设备故障智能分类模型,并以某铁路局 2012~2016 年铁路信号设备故障文本数据验证了模型的有效性。李新琴等学者[10]针对铁路安全事故隐患文本数据分类提出了进化集成分类器模型,并通过对某铁路局供电接触网安全事故隐患文本数据实验分析证明了进化集成分类器模型优于单个决策树分类器和 Bagging 集成分类器分类结果。周庆华等学者[11]针对铁路信号设备故障短文本数据提出一种基于 Word2Vec+MCNN 的文本挖掘分类方法,并以实验证明了该方法可以更好地达到分类效果,具有较高的分类准确率。韩广等学者[12]针对铁路行车事故文本具有专业词多、描述文本长短不一的特点,提出一种结合双通道双向长短时记忆网络和注意力机制的铁路行车事故文本等级分类方法,并证明了该方法的有效性。但针对接触网异物文本数据的研究还不多见,以全路接触网异物信息文本为基础,综合考虑算法准确率和时间消耗对比多个常用的多分类模型,提出了基于梯度提升决策树(Gradient Boosting Decision Tree,GBDT)模型的接触网异物分类方法,并根据各铁路局集团公司实际接触网异物文本数据进行实验分析,结果表明基于 GBDT 的接触网异物分类模型优于相比较的常用分类模型。以文本挖掘技术分析接触网异物并进行准确高效分类,既能对接触网异物处理做出及时预判,又能为管理人员在信息化管理接触网异物信息方面提供便利,这对正在铁路行业进行统型建设的铁路供电管理信息系统有着重要的意义。

## 1 数据预处理

### 1.1 文本分词及清洗

在中文文本分类中,领域专有词典可快速高效地提高中文分词效果,支撑关键特征提取。针对接触网异物文本中专有词汇,建立“接触网异物词典”,以 jieba 分词工具进行加载后对文本数据进行分词,对分词后的文档去除停用词,如“啊”、“吧”等,并删除文档中的数字、中英文字符等无效内容,提高文本特征质量。文本分词及清洗后部分数据见表 1。

**表 1 分词及清洗后部分数据**

**Table 1 Partial data after word segmentation and cleaning**

| 类别 | 文本 |
|---|---|
| 编织袋 | 司机 大步 站 十 三间房 站间 上行 线处 接触网 搭挂 异物 编织袋 降弓 通知 哈密 供电段 分 哈密 现场 检查 异物 销记 影响 货车 列 |
| 风筝 | 司机 报 京哈 高速 上行 线 接触网 挂 异物 影响 行车 需降 弓 时 分 现场 人员 报 巡视 发现 京哈 高速 上行 线 处 支柱 哈侧 吊 承力 索侧 挂 异物 风筝 利用 间接 带电 处理完毕 影响 客车 列降 弓 |
| 鸟窝 | 唐包线 古 营盘 接触网 工区 汇报 巡视 发现 庙 梁西 土城 上下行 线 处供 供 支柱 鸟巢 点 分钟 列调 点庙 梁 变电所 停电 处理完毕 恢复 供电 影响 货车 列 |

以接触网异物数据中树枝类别数据为例,分词及清洗后以词云图的方式进行可视化,其结果如图 1 所示。

### 1.2 文本表示

文本表示在自然语言处理流程中有着承上启下的作用,用于将文本转化成计算机可以识别的数字

化形式[13]。词频－逆文档频率(Term frequency－Inverse document frequency，TF－IDF)算法主要用于挖掘文本中的关键词，可评估前文选择的词在文本中是否具有代表性[14]。因此，使用 TF－IDF 算法实现接触网异物文本关键词提取和向量化表示。TF－IDF 算法由词频(Term Frequency，*TF*) 和逆文档频率(Inverse Document Frequency，*IDF*) 两部分组成，可被描述为式(1)、式(2)：

$$TF = \frac{\text{关键词在文档中出现的次数}}{\text{文档中词总数}} \tag{1}$$

$$IDF = \log\left(\frac{\text{语料库中文档总数}}{\text{含关键词的文档数} + 1}\right) \tag{2}$$

以 TF－IDF 算法将文本特征转为计算机可识别的数字化形式，文本数字化表示部分结果见表 2。


**图 1　接触网异物树枝类别词云图**

**Fig. 1　Categorical word cloud map of foreign objects in the overhead contact system**

**表 2　文本向量化表示部分结果**

**Table 2　Partial results of text vectorization representation**

| 特征词 | 数据表示 | 特征词 | 数据表示 | 特征词 | 数据表示 | 特征词 | 数据表示 |
|---|---|---|---|---|---|---|---|
| 接触网 | 6 221 | 风筝 | 13 526 | 绳索 | 10 463 | 抢修 | 5 949 |
| 编织袋 | 10 493 | 鸟窝 | 13 781 | 绳子 | 10 455 | 晚点 | 6 771 |
| 塑料袋 | 3 754 | 氢气球 | 8 071 | 树枝 | 7 600 | 调度 | 11 825 |
| 尼龙袋 | 4 682 | 气球 | 8 067 | 雨衣 | 13 307 | 异物 | 5 361 |
| 塑料布 | 3 745 | 尼龙绳 | 4 680 | 异常 | 5 359 | 停电 | 1 219 |
| 塑料薄膜 | 3 753 | 线绳 | 10 308 | 杂物 | 7 018 | 正线 | 7 933 |

## 2　基于 GBDT 模型的接触网异物分类

### 2.1　分类回归树模型

分类回归树模型(Classification and Regression Tree，CART) 由 Breiman 等学者在 1984 年提出[15]，是一种二叉树模型，由特征选择、决策树生成以及决策树剪枝组成，既可以用于分类、也可以用于回归，直接用于分类效果要优于回归。CART 模型结构如图 2 所示。

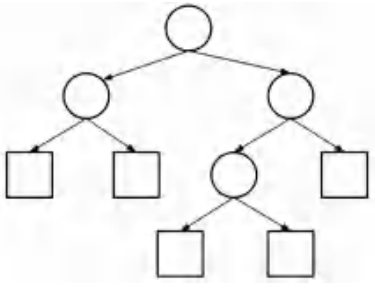


**图 2　CART 模型结构示意图**

**Fig. 2　Schematic diagram of CART model structure**

#### 2.1.1　特征选择

CART 模型基于基尼(Gini)系数最小化准则来进行特征选择。*Gini* 系数是指模型的不纯度，基尼系数越小，则不纯度越低，特征越好。*Gini* 系数表达式为：

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{3}$$

其中，$D$ 表示样本；$K$ 表示类别数；$p_k$ 表第 $k$ 个类别的概率。

对于每一个特征值 $A$，可能取值 $a$，将样本数据集 $D$ 分割为 $D_1$ 和 $D_2$ 两个子集，则样本数据集 $D$ 对于特征 $A$ 的 *Gini* 系数可由式(4) 来求得：

$$Gini(D,A) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2) \tag{4}$$

#### 2.1.2　决策树生成

CART 模型通过最小化样本数据集中每个特征可能的取值的 *Gini* 系数生成决策树，生成算法描述如下。

**算法 1　CART 模型决策树生成**

**输入**　训练集 $D$，Gini 系数阈值，样本个数阈值

**输出**　决策树

1.计算现有特征对训练集 $D$ 的 Gini 系数。对每一个特征 $A$，判定每一个可能值 $a$ 的正确与否，记为“是”或“否”，将训练集 $D$ 划分为 $D_1$ 和 $D_2$ 两个子集，利用式(4)计算 $A = a$ 的 Gini 系数；

2.在特征 $A$ 和可能的取值 $a$ 中选取 Gini 系数最

小的作为最优特征和最优切分点,根据该最优特征和最优切分点生成 2 个子节点,将训练集 $D$ 的特征划分至 2 个子节点;

3.对 2 个新划分的子节点递归步骤 1、2,直至满足阈值条件;

4.生成决策树。

2.1.3 决策树剪枝

CART 模型决策树剪枝算法是对于生成的决策树剪去一些子树,使得决策树更简单,防止产生过拟合现象。剪枝算法描述如下。

**算法 2 CART 模型决策树剪枝**

**输入** CART 模型生成的决策树 $T_0$

**输出** 最优决策子树 $T_\alpha$

1.初始化:$k=0$,$T=T_0$,设 $\alpha=+\infty$;

2.自叶子节点开始自下而上地对内部节点 $t$ 计算 $C(T_t)$、$|T_t|$、$g(t)=\dfrac{C(t)-C(T_t)}{|T_t|-1}$、$\alpha=\min(\alpha, g(t))$。

其中,$T_t$ 表示以 $t$ 节点为根节点的子树,$C(T_t)$ 表示训练集的预测误差,$|T_t|$ 表示以 $t$ 节点为根节点的子树的叶子节点数;

3.对 $\alpha=g(t)$ 的内部节点 $t$ 剪枝,以多数表决法确定剪枝后节点 $t$ 的类别,得到树 $T^{k+1}$,$k=k+1$。

4.递归步骤 2、3,直至 $T^k$ 为一个三节点树。

5.采用交叉验证法求出得到的子树($T_0,T_1,\cdots,T_n$)中的最优子树 $T_\alpha$。

## 2.2 GBDT 模型原理

GBDT 模型于 2001 年由 Friedman 提出[16],是一种以 CART 回归树为基学习器的梯度提升算法,以选定的数据集为基础,通过构造多组基学习器(CART 回归树),在充分考虑每组基学习器的权重下把多组基学习器的结果累加起来作为最终的预测输出。其模型具有稳定性好、运算时间少、算法参数少的优点[17],训练原理如图 3 所示。

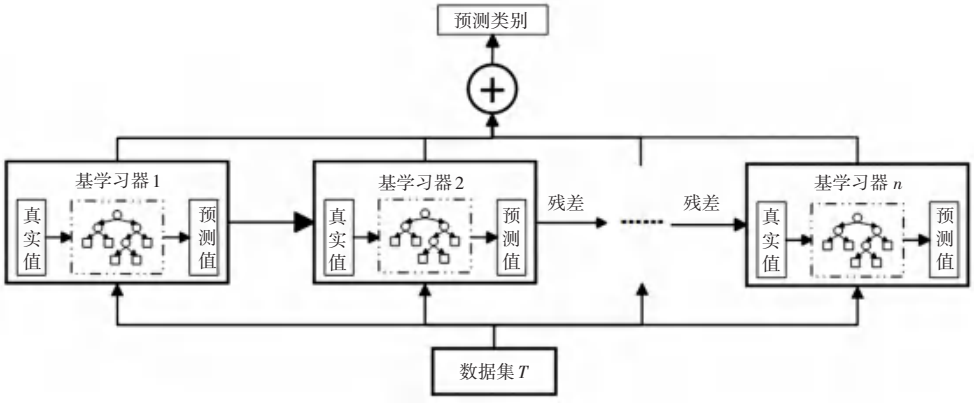



**图 3 GBDT 模型训练原理**

**Fig. 3 Training principles of GBDT model**

GBDT 模型同时是一种基于 Boosting 算法的集成思想的加法模型。Boosting 主要表现在通过构建一系列的基学习器来提高算法分类的准确率[18]。基于 Boosting,在残差减小的梯度方向上建立新的决策树[19]。此处需用到的公式为:

$$F_M(x)=\sum_{m=1}^{M}T(x;\Phi_m) \tag{5}$$

其中,$T(x;\Phi_m)$ 表示决策树;$\Phi_m$ 表示决策树参数;$M$ 表示决策树个数。

GBDT 模型通过最小化损失函数计算第 $m$ 棵决策树的参数,函数表达式可写为:

$$\Phi_m=\arg\min\sum_{i=1}^{N}L(y_i,T_{m-1}(x_i)+T(x_i;\Phi_m)) \tag{6}$$

其中,$L$ 表示 GBDT 模型的损失函数,采用的对数似然损失函数,可由式(7)进行描述:

$$L(y,F(x))=-\sum_{k=1}^{K}y_k\log(p_k(x)) \tag{7}$$

其中,$K$ 表示类别数,$p_k(x)$ 表示第 $k$ 类的概率。

## 2.3 基于 GBDT 模型的接触网异物分类

基于 GBDT 模型的接触网异物分类流程主要包含选择数据集、数据预处理、数据拆分、模型训练和评估。接触网异物分类数据集主要包含编织袋、风筝、鸟窝、气球、树枝、塑料布、塑料袋和线网绳八个类别数据,其中由于数据量原因,线网绳类别数据由原始数据中线索(绳索)和尼龙网(绳)两个类别数据共同组成。数据预处理主要是利用 Jieba 分词工具加载“供电异物词典”针对接触网异物分类原始数据集进行分词,对分词后文本开展数据清洗,最后利用 TF-IDF 算法实现文本关键词抽取及向量化表

示。针对预处理后的数据集按照 8∶2 的比例随机抽取训练集、测试集数据，用来训练和验证模型的性能优劣。以训练集数据进行 GBDT 模型训练，测试集数据验证模型优良，并统计测试集数据真实类别和预测类别数量用以对训练模型的评估。基于 GBDT 模型的接触网异物分类流程如图 4 所示。

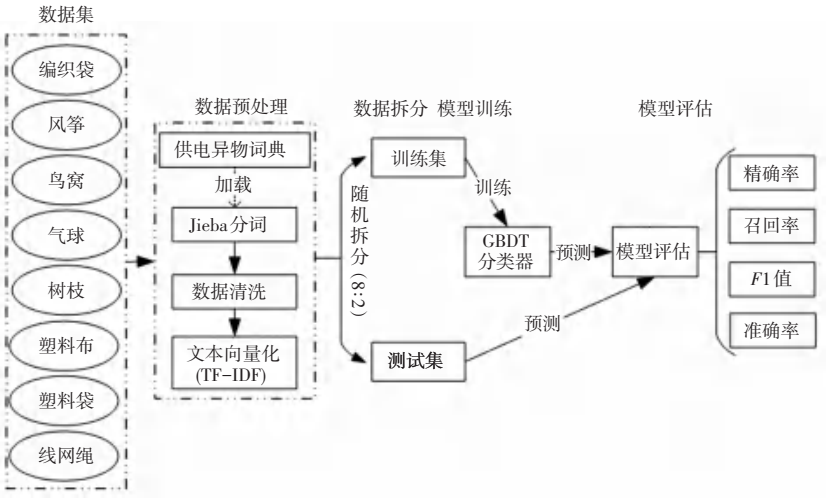



**图 4　基于 GBDT 模型的接触网异物分类**

**Fig. 4　Classification of foreign objects in contact system based on GBDT model**

为验证所选择模型的好坏，需要采用对应的模型评估指标，由图 4 可以观察到，基于 GBDT 的接触网异物分类模型的性能可采用精确率（*Precision*）、召回率（*Recall*）、*F*1 值和准确率（*Accuracy*）四个指标进行评估，每个评估指标数学含义具体如下。

（1）精确率。又叫做查准率，是针对预测结果而言的，其含义是在被所有预测为正的样本中实际为正样本的概率，研究推得的数学公式为：

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

（2）召回率。又叫查全率，是针对原样本而言的，其含义是在实际为正的样本中被预测为正样本的概率，研究推得的数学公式为：

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

（3）*F*1 值。同时考虑精确率和召回率，使两者达到平衡，研究推得的数学公式为：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

（4）准确率。表示预测正确的结果占总样本的百分比，研究推得的数学公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

式（8）~（11）中，*TP* 表示实际样本为正样本且预测为正样本；*TN* 表示实际样本为负样本且预测为负样本；*FP* 表示实际样本为负样本但预测为正样本；*FN* 表示实际样本为正样本但预测为负样本。接触网异物分类模型的性能优劣最终由模型准确率来表示。

## 3　实验分析

### 3.1　实验数据

以 2016 年 1 月至 2023 年 4 月全国铁路接触网异物文本数据共计 14 470 条有效数据进行实验，按照 8∶2 的比例将各类别接触网异物文本数据进行自动划分训练集和测试集，划分后训练集 11 580 条，测试集 2 890 条。各类别训练集和测试集实验数据见表 3。由表 3 可以观察到，塑料布类别的数据最多，塑料袋数据次之，且二者数据之和占总数量的 1/2 以上，可见接触网常见的异物主要是塑料布和塑料袋。

**表 3　各类别训练集和测试集数据量**

**Table 3　Data volume of training sets and testing sets for each category**

| 数据集 | 编织袋 | 风筝 | 鸟窝 | 气球 | 树枝 | 塑料布 | 塑料袋 | 线网绳 |
|---|---|---|---|---|---|---|---|---|
| 训练集 | 328 | 863 | 1 946 | 1 139 | 552 | 3 468 | 2 802 | 482 |
| 测试集 | 81 | 215 | 486 | 284 | 138 | 866 | 700 | 120 |

### 3.2 对比模型

基于 GBDT 的接触网异物分类模型选择了 K 最近邻(K-Nearest Neighbor ,KNN)[20]、多项式朴素贝叶斯(Multinomial Naive Bayes ,MNB)[21]、逻辑回归(Logistic Regression ,LR)[22]、随机森林(Random Forest ,RF)[23]、决策树((Decision Tree, DT)[24]、Adaptive Boosting ( Adaboost )[25]、支持向量机(Support Vector Machine ,SVM)[26]共 7 个分类模型进行实验对比,每个模型的含义如下。

KNN 模型核心思想是给定度量距离的方式,判断一个样本在特征空间中的 $k$ 个最相邻的样本中是否大多数属于某一个类别。若是,则判定该样本也属于这个类别;MNB 模型是指在假设特征条件独立且文档中词出现的位置也互不影响下,将重复出现的词视为多次出现,计算每个类别在训练样本中出现的概率和每个特征在某类别文档出现次数相对于某类别文档中所有词总数的条件概率;LR 模型多分类的核心思想是将多分类问题转换为二分类问题,即每次将其中一个类别和剩余类别看成一类;RF 模型是利用多个决策树来训练样本并进行预测的一种分类器,可以解决一个决策树分类的误差问题和过拟合问题;DT 模型核心思想是基于树结构从根节点开始测试待分类项中的特征属性,按照特征进入相应分支,直至达到叶子节点并输出决策结果;Adaboost 模型核心思想是针对同一数据集训练弱分类器,通过调整错误样本的权重迭代升级分类器,最终形成一种强分类器作为决策分类器;SVM 模型核心思想是通过构建一个“超平面”,利用“超平面”进行分类。

### 3.3 模型调参

以机器学习工具包 sklearn 中的 GBDT 多分类模型进行实验分析,为达到最佳效果需调整模型相关参数,主要针对模型决策树数量、深度参数进行调整,同时选择合适的评估指标平均值计算方式。

#### 3.3.1 评估指标平均值计算方式

多分类评估指标平均值计算方式有 weighted、macro、micro 三种,以默认参数值下的 GBDT 模型进行训练,选用精确率、召回率和 $F1$ 值进行评价,其评估结果见表 4。由表 4 可以观察到,macro 计算的精确率、召回率和 $F1$ 值相对于 weighted 和 micro 较低,而 micro 计算出的精确率、召回率和 $F1$ 值三者相等,不利于模型测试结果比较。因此,选择 weighted 方式计算评估指标平均值。

**表 4 不同评估指标平均值计算方式评估结果**

**Table 4 Evaluation results of different calculation methods for average values of evaluation indicators**

| 评估指标平均值计算方式 | 精确率 | 召回率 | $F1$ 值 |
|---|---|---|---|
| weighted | 94.26 | 94.39 | 94.15 |
| macro | 91.90 | 89.46 | 90.19 |
| micro | 94.33 | 94.33 | 94.33 |

#### 3.3.2 决策树深度

在选择了 weighted 评估指标平均值计算方式后,根据决策树深度探索模型评估结果如图 5 所示。由图 5 可以观察到,决策树深度为 4 时模型效果较好。因此,选择决策树深度为 4 进行模型训练。

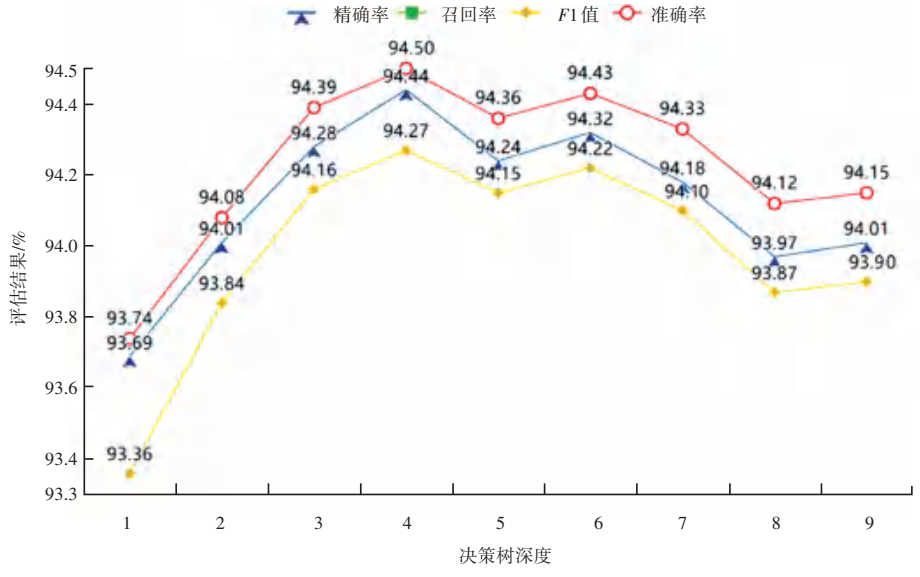



**图 5 随决策树数深度增加评估结果变化**

**Fig. 5 Changes in evaluation results as the number of decision trees increases in depth**

3.3.3　决策树个数

GBDT 模型默认决策树个数为 100，以 weighted 评估指标平均值计算方式，决策树深度为 4 为基础，以 100 为中间值，探索 50～150 范围内随决策树增加评估结果变化情况，结果如图 6 所示。由于决策树个数太小会导致模型欠拟合，太大会导致模型过拟合，结合图 6 的评估结果，训练模型决策树个数选择 120。

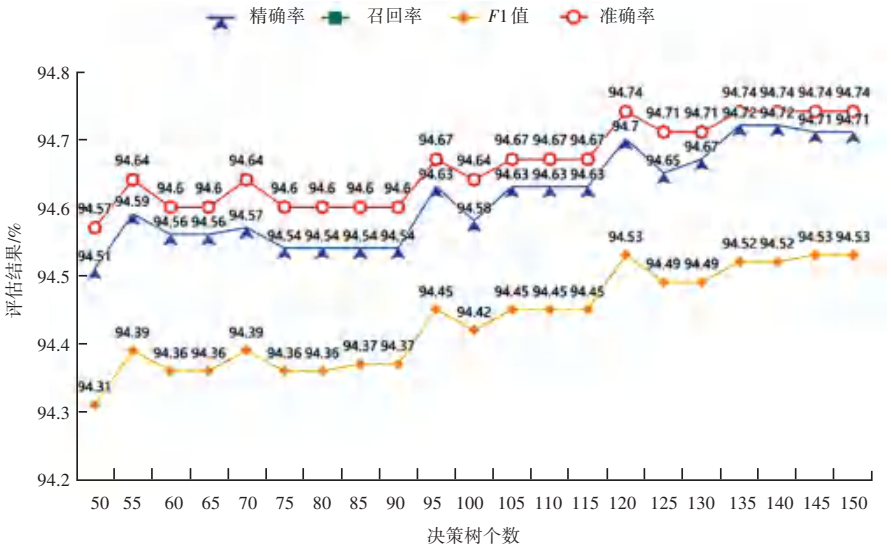



**图 6　随决策树数个数增加评估结果变化**

**Fig. 6　Changes in evaluation results as the number of decision trees increases**

## 3.4　实验结果

基于 TF-IDF 算法，以 weighted 方式计算评估指标平均值，决策树深度为 4，决策树个数为 120 进行 GBDT 模型训练和测试，同时与 KNN、MNB、LR、RF、DT、Adaboost、SVM 共 7 个分类模型进行实验对比，各模型实验结果见表 5。

**表 5　各模型实验结果对比**

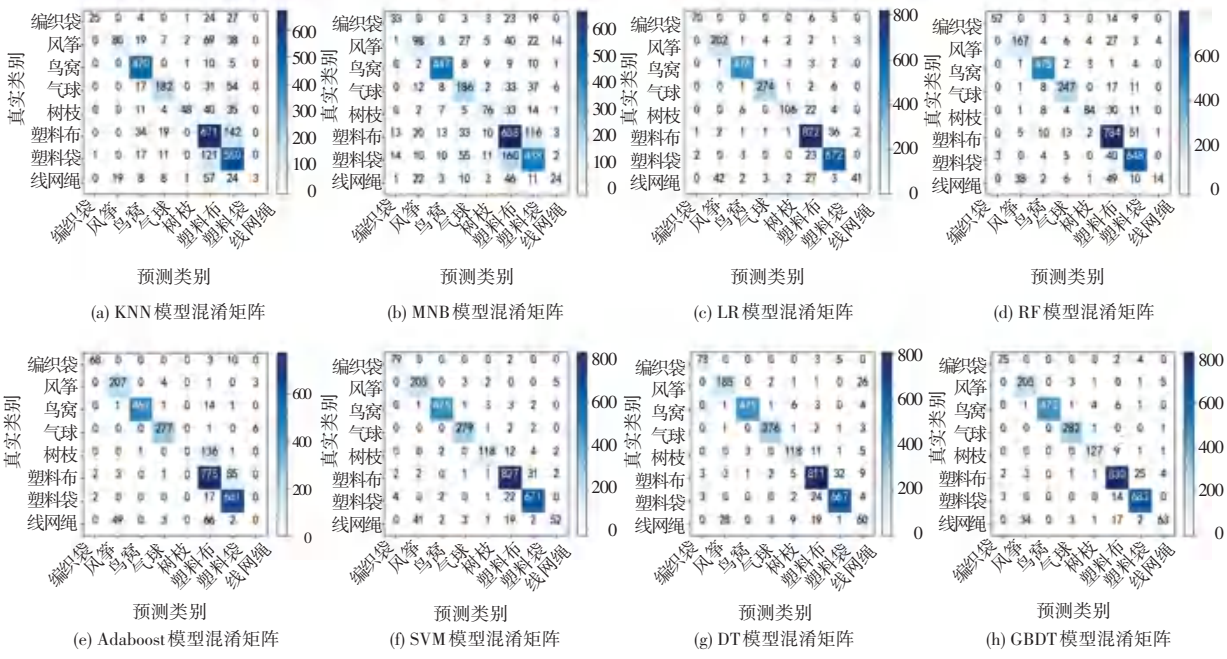
**Table 5　Comparison of experimental results of various models　%**

| 模型 | 精确率 | 召回率 | *F*1 值 | 准确率 |
|---|---|---|---|---|
| TF-IDF+KNN | 73.43 | 70.21 | 67.68 | 70.21 |
| TF-IDF+MNB | 67.22 | 67.82 | 67.04 | 67.82 |
| TF-IDF+LR | 92.21 | 92.15 | 91.57 | 92.15 |
| TF-IDF+RF | 85.27 | 85.50 | 84.13 | 85.50 |
| TF-IDF+Adaboost | 78.94 | 85.71 | 82.04 | 85.71 |
| TF-IDF+SVM | 93.63 | 93.67 | 93.32 | 93.67 |
| TF-IDF+DT | 91.97 | 92.08 | 92.01 | 92.08 |
| TF-IDF+GBDT | 94.70 | 94.74 | 94.53 | 94.74 |

针对表 5 各模型实验结果可以观察到，GBDT 模型的评估指标，即精确率、召回率、*F*1 值和准确率均优于对比模型，其准确率达到了 94.74%。相对于对比模型中仅次于 GBDT 模型训练效果的 SVM 模型的训练结果，准确率提高了 1.07%。考虑模型分类效果，同时对比分析表 5 中各类别训练集和测试集数据量，可知 GBDT 模型集成多个弱分类器模型在文本分类效果上具有一定的优越性，能够很好地应对文本分类数据不平衡的问题。

各模型在接触网异物各类别的分类结果以混淆矩阵表示，实验结果混淆矩阵如图 7 所示。图 7 中，混淆矩阵中列表示预测值，列的总数表示预测为该类别的数量；行表示真实值，行的总数表示该类别的真实数量。

GBDT 模型测试集各类别实验结果见表 6。由表 6 结合图 7 可以观察到，除了塑料袋类别，各分类模型在其他 7 个类别的数据上的评估指标精确率、召回率和 *F*1 值均取得了良好的效果。通过对数据进行分析发现，塑料袋类别的评估指标效果较差主要原因有 2 个。一是和塑料布类别数据关键特征有许多交叉，二是由于原始数据由各铁路局集团公司的不同人员人工分类上传，塑料布类别和塑料袋类别数据误分较多。结合图 7 各模型混淆矩阵中塑料布类别和塑料袋类别预测错误的数据结果，可以观察到塑料布类别数据预测错误结果多为塑料袋类别，而塑料袋类别数据预测错误结果多为塑料布类别，这是由原始数据特征导致，同时也证明了研究基于文本挖掘算法的接触网异物分类方法的必要性。




(a) KNN 模型混淆矩阵 (b) MNB 模型混淆矩阵 (c) LR 模型混淆矩阵 (d) RF 模型混淆矩阵

(e) Adaboost 模型混淆矩阵 (f) SVM 模型混淆矩阵 (g) DT 模型混淆矩阵 (h) GBDT 模型混淆矩阵

**图 7 各模型混淆矩阵**

**Fig. 7 Confusion matrix of each model**

**表 6 GBDT 模型各类别实验结果**

**Table 6 Experimental results of various categories of GBDT model** %

| 模型 | 精确率 | 召回率 | *F*1 值 |
|---|---|---|---|
| 编织袋 | 90.36 | 92.59 | 91.46 |
| 风筝 | 84.36 | 95.35 | 89.52 |
| 鸟窝 | 100.00 | 97.33 | 98.64 |
| 气球 | 97.24 | 99.30 | 98.26 |
| 树枝 | 94.07 | 92.03 | 93.04 |
| 塑料布 | 93.05 | 95.84 | 94.43 |
| 塑料袋 | 48.93 | 97.57 | 65.17 |
| 线网绳 | 85.14 | 83.54 | 84.33 |

## 4 结束语

当前,国内电气化铁路里程和高铁里程稳居世界第一,接触网作为电气化铁路的重要组成部分,快速识别和防止接触网异物入侵直接关系到铁路的正常运行[27-28]。结合全国铁路接触网异物入侵文本数据,通过对接触网异物分类进行研究,得到的结论如下:

(1)首次对接触网异物文本数据进行研究分析,填补了接触网异物领域文本数据研究的空白。

(2)以 2016 年 1 月至 2023 年 4 月全国铁路接触网异物文本数据为基础,提出了基于 GBDT 模型的接触网异物分类方法,同时对比分析了 KNN、MNB、LR、RF、DT、Adaboost、SVM 共 7 个多类别文本分类算法,并以实验证明了基于 GBDT 模型的接触网异物分类方法的优越性。

(3)由于铁路接触网异物文本数据不均衡,实验结果证明,基于 GBDT 模型的接触网异物分类方法能够有效地应对文本分类数据不均衡问题并取得很好的分类效果,且具备可靠性和可行性,有一定的推广和应用价值。

(4)由各类别数据实验结果可知,塑料袋类别数据识别效果不理想,通过数据及实验分析,建议重新考虑和划分铁路接触网异物文本数据类别,保证各类别数据尽可能地多,且最大限度降低各类别的文本相似性。

## 参考文献

[1] 盛望群. 基于 CDEGS 的交流电气化铁路对沿线油气管道电磁干扰影响研究[J]. 铁道科学与工程学报,2020,17(8):2101-2108.

[2] 杨忠平. 既有铁路电气化改造工程环境保护对策[J]. 铁路节能环保与安全卫生,2021,11(6):29-31.

[3] 蒋欣兰,贾文博. 高铁接触网异物侵入的机器视觉检测方法[J]. 计算机工程与应用,2019,55(22):250-257.

[4] 白康康. 大数据网络安全防御中人工智能技术的应用分析[J]. 计算机应用文摘,2022(2):38-40,46.

[5] 缪炳荣,张卫华,刘建新,等. 工业 4.0 下智能铁路前沿技术问题综述[J]. 交通运输工程学报,2021,21(1):115-131.

[6] 徐伟,吴泽彬,刘建新,等. 高铁接触网异物自动化智能检测方法[J]. 中国铁路,2019(10):39-44.

[7] 王科理,高福来,杨鹏,等. 基于深度学习的接触网鸟巢异物识别研究[J]. 铁道机车车辆,2022,42(2):116-121.

[8] TAN Ping, LI Xufeng, DING Jin, et al. Mask R-CNN and multifeature clustering model for catenary insulator recognition and defect detection[J]. Journal of Zhejiang University-Science A (Applied Physics & Engineering),2022,23(9):745-757.
[9] 杨连报,李平,薛蕊,等. 基于不平衡文本数据挖掘的铁路信号设备故障智能分类[J]. 铁道学报,2018,40(2):59-66.
[10]李新琴,史天运,李平,等. 基于进化集成分类器的铁路安全隐患智能分类[J]. 交通信息与安全,2019,37(2):33-39.
[11]周庆华,李晓丽. 基于 MCNN 的铁路信号设备故障短文本分类方法研究[J]. 铁道科学与工程学报,2019,16(11):2859-2865.
[12]韩广,卜桐,王明明,等. 基于双通道双向长短时记忆网络的铁路行车事故文本分类[J]. 铁道学报,2021,43(9):71-79.
[13]赵京胜,宋梦雪,高祥,等. 自然语言处理中的文本表示研究[J]. 软件学报,2022,33(1):102-128.
[14]王心仪,程剑锋,刘育君. 基于 TF-IDF 加权朴素贝叶斯算法的 ATP 车载设备测试案例分类研究[J]. 铁路计算机应用,2022,31(12):8-12.
[15] BREIMAN L, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression trees (CART)[J]. Biometrics, 1984, 40(3):358.
[16] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.
[17]廖璐,张亚东,葛晓程,等. 基于 GBDT 的列车晚点时长预测模型研究[J]. 铁道标准设计,2021,65(8):149-154,176.
[18]张远旗. 分布式 Boosting 算法研究及其在图像目标检测中的应用[D]. 成都:电子科技大学,2020.
[19]钟敏慧,张婉露,李有儒,等. 基于 GBDT 的铁路事故类型预测及成因分析[J]. 自动化学报,2022,48(2):470-478.
[20]COVER T M, HART P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory,1967,13(1):21-27.
[21]贺鸣,孙建军,成颖. 基于朴素贝叶斯的文本分类研究综述[J]. 情报科学,2016,34(7):147-154.
[22]王济川,郭志刚. Logistic 回归模型方法及应用[M]. 北京:高等教育出版社,2001.
[23]BREIMAN L. Random forest[J]. Machine Learning,2001,45:5-32.
[24]RUTKOWSKI L, PIETRUCZUK L, DUDA P, et al. Decision trees for mining data streams based on the gaussian approximation[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(6):1272-1279.
[25]董乐红,耿国华,高原. Boosting 算法综述[J]. 计算机应用与软件,2006,23(8):27-29.
[26]CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.
[27]许建国,李志锋,刘杰. 用奋斗铸就现代化铁路的动力之源—中国电气化铁路建设十万公里纪实[EB/OL]. [2020-12-21]. https://finance.ifeng.com/c/82No7DRF11B.
[28]池瑞,邱国龙,曾庆森,等. 高速铁路接触网系统维修决策优化[J]. 铁道科学与工程学报,2023,20(1):53-62.