第53卷第6期 南京航空航天大学学报 Vol. 53 No. 6
2021年12月 Journal of Nanjing University of Aeronautics & Astronautics Dec. 2021

DOI:10.16356/j.1005-2615.2021.06.003

基于LightGBM的航班延误多分类预测

丁建立,孙 玥
(中国民航大学计算机科学与技术学院,天津 300300)

摘要: 航班延误是民航业的一大难题,提前对航班的延误情况进行预测,以采取合理的应对措施,对缓解航班延误产生的负面影响有着重要意义。为提升预测性能,提出一种基于轻量级梯度提升机(Light gradient boosting machine,LightGBM)的航班延误多分类预测模型。该模型结合航班信息与天气信息,运用方差过滤与递归特征消除进行特征筛选,并采用合成少数过采样技术(Synthetic minority oversampling technique,SMOTE)与Tomek Link对数据进行不平衡处理,最后使用LightGBM进行建模,实现对航班延误时长的多分类预测。为验证模型的合理性,将所提模型与其他先进算法构建的模型进行对比。实验结果表明,所提模型在各种预测性能指标上结果更优,将预测精度提升至90%以上,同时大幅度降低了训练时间成本。

关键词: 航班延误;预测模型;轻量级梯度提升机;贝叶斯调参

中图分类号: TP181 **文献标志码:** A **文章编号:** 1005-2615(2021)06-0847-08

Multi-classification Prediction of Flight Delay Based on LightGBM

DING Jianli, *SUN Yue*
(College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Delay of flights is a core problem in civil aviation industry. It is significant to predict flight delay situation in advance so as to take reasonable measures to negative effects of flight delay. In order to improve the prediction performance, a flight delay multi-classification prediction model based on light gradient boosting machine (LightGBM) is put forward in this paper. This model can screen features by using variance filtering and recursive feature elimination according to flight information and weather information. It uses synthetic minority oversampling technique (SMOTE) and Tomek Link to deal with unbalanced data. Finally, LightGBM is used to build models, and multi-classification prediction of flight delay lengths can be realized. In order to verify the rationality of the model, this paper compares the proposed model with models constructed by other advanced algorithms. The experimental results show that the proposed model performs better in terms of various prediction performance indexes and can improve the prediction accuracy to 90% or higher. The model can also greatly reduce the training time and cost.

Key words: flight delay; prediction model; light gradient boosting machine (LightGBM); Bayesian optimization

由于恶劣天气、空域限制等诸多因素,航班延误率居高不下。据民航局发布《2019年民航行业发展统计公报》所示,中国客运航空公司的全年平均正常航班率为81.65%,有近20%的航班发生了延误现象。航班延误的频繁发生,不仅会影响机场以及管制部门的正常运行,额外增加航空公司的运营成本,造成公共运输服务资源的浪费,还会影响旅客的出行体验。在发生大面积延误时,大量滞留在机场的旅客很可能会引发混乱与纠纷,甚至与工作人员发生冲突,危害社会秩序与安全。因此,对

基金项目: 国家自然科学基金民航联合基金重点(U2033205)资助项目。
收稿日期: 2021-05-10;**修订日期:** 2021-10-19
通信作者: 孙玥,女,硕士研究生,E-mail:2019051012@cauc.edu.cn。

引用格式: 丁建立,孙玥. 基于LightGBM的航班延误多分类预测[J]. 南京航空航天大学学报,2021,53(6):847-854.
DING Jianli, SUN Yue. Multi-classification prediction of flight delay based on LightGBM[J]. Journal of Nanjing University of Aeronautics & Astronautics,2021, 53(6):847-854.

航班的延误情况提前进行预测，相关负责人员可以根据预测不同的延误程度，提前进行有序的调度与合理的资源分配，防止因延误的累积性造成恶性循环，旅客也可以对延误情况有一定的心理准备，及时对自己的行程安排进行调整，尽可能地减少航班延误带来的不利影响。

在航班延误预测的相关工作中，较为传统的预测方法主要包括回归分析、贝叶斯网络以及支持向量机等。Klein等[1]通过天气影响交通指数(Weather impact traffic index，WITI)建立了一个机场延误预测的多元回归模型，对全年对流和非对流天气下航班延误的预测性能进行比较；文献[2]基于贝叶斯网络方法，通过建立预测阶段和可靠性阶段两个步骤，提出了一种预测和评估机场到达系统运行状态的方法，可以得出对机场性能有影响的因素之间的相互依赖关系；Álvaro等[3]采用贝叶斯网络方法对机场到达航班的拥挤程度和延误情况进行预测，并利用马尔可夫链技术对多状态系统进行可靠性分析；文献[4]建立了一种改进的支持向量机模型，采用主成分分析法进行特征筛选，并利用航班计划的周期性，将历史数据中的离港航班数和离港延误率作为先验知识提供给支持向量机，提高模型的精度。

随着社会信息化程度的迅速发展，数据集的规模与复杂程度也在不断提升，研究更多地运用了数据挖掘的方法。Achenbach等[5]将线性回归和梯度提升进行结合，提出了一种短途航班到达时间预测和成本指数优化模型，考虑了3种不同的飞行距离来模拟成本指数变化对登机口到达时间的影响；Gui等[6]分别训练了两种预测模型，并从分类与回归任务两个角度比较了模型的性能；周洁敏等[7]利用随机森林进行特征筛选，建立了弹性神经网络预测模型，对航班落地延误时间进行预测；吴仁彪等[8]基于DenseNet模型构建航班延误预测模型，解决了深层训练时的梯度消失现象，并提出SE-DenseNet模型，实现了特征提取过程中的权重自适应标定，减少了信息冗余的问题。

在航班延误的预测问题上，相关研究已经可以获得较高的准确率。本文进一步对影响航班延误的因素进行完善，增加天气因素与前序航班相关因素，并针对训练时间较长的问题进行改进，选用运行速度快、占用内存低的轻量级梯度提升机(Light gradient boosting machine，LightGBM)算法[9]进行建模。

LightGBM是一种分布式的梯度Boosting框架，目前的相关研究已广泛涉及医学[10-11]、机械故障检测[12]和风力发电功率预测[13]等多个方向。本文提出一种基于LightGBM的航班延误预测多分类模型，可以实现对航班延误时长的多等级预测，达到更快的训练速度与更好的预测性能。

1 数据及处理方法

1.1 数据来源

本文选定的目标机场为纽瓦克自由国际机场，，其航班的历史数据来源于美国交通运输统计局，该机构统计了从1987年至今的美国全空域航班信息。选取的数据为2019全年历史航班数据，内容主要包括时间信息、航空公司信息、机场信息与延误情况等共120个特征。

天气的历史数据来源于美国国家海洋和大气管理局，选取的天气数据同样以纽瓦克自由国际机场为目标，且与选定的航班数据有着相同的时间跨度，便于后续的数据融合。天气数据包括测定时间、测定地点、气温、露点、降水、相对湿度、云层状况、能见度、风向、风速和异常天气类型等共29个特征。

1.2 数据预处理

1.2.1 缺失值处理

将航班数据与天气数据根据时间标签进行融合，最终形成统一的数据集，共有310 447条数据。通过对数据集进行检查，发现部分特征的缺失率较高，共有59个特征缺失率达到了7成以上。对于这些特征，采取直接删除的方式。余下特征的缺失率在5%以下，可以根据不同特征的特点进行填充。部分特征处理方式如表1所示。在“降水量”这一特征中，降水量为0的样本占所有样本的74.4%，可以直接用众数0对其缺失值进行填充；“云层状况”这一特征在相邻时刻内的变化不会过大，可以用上一个时刻的值来对缺失时刻的值进行填充。经过缺失值处理后的数据集共包含85个特征。

表1 特征缺失值处理方式(部分)

Table 1 Treatment for certain characteristic missing value (Part)

特征	缺失率/%	处理方法
CancellationCode	96.63	直接删除
WindGustSpeed	81.97	直接删除
CarrierDelay	71.62	直接删除
PriorArrDelay	16.42	以0填充
Precipitation	0.73	以众数填充
SkyConditions	0.50	以上一时刻的值填充
⋮	⋮	⋮

1.2.2　前序航班

同一架飞机在一天中会执行多个连续航班的任务，如果前序航班发生到达延误，当前航班的离港时间也会受到延误波及。因此，本模型将航班信息按照飞机尾翼号划分成组，并按照时间排序，找出当前航班的前序航班相关信息。将前序航班预计到达时间、实际到达时间以及延误时间这3个特征作为数据集的新特征，充分考虑前序航班的延误情况对当前航班的影响。

如表2所示的4个航班，尾翼号为“234NV”，其中序号1与序号3为当日第1班航班，没有前序航班，故相关特征为空；序号2的前序航班实际到达时间早于预计到达时间，故延误时间为0；序号4的前序航班实际到达时间晚于预计到达时间10 min，故延误时间为10 min。

表2　前序航班处理方式（部分）

Table 2　Pre-order flight processing method (Part)

序号	尾翼号	计划到达时间	实际到达时间	前序航班预计到达时间	前序航班实际到达时间	延误时间/min
1	234NV	2019-04-06 11:00:00	2019-04-06 10:24:00			
2	234NV	2019-04-06 14:00:00	2019-04-06 13:39:00	2019-04-06 11:00:00	2019-04-06 10:24:00	0
3	234NV	2019-09-05 10:00:00	2019-09-05 10:10:00			
4	234NV	2019-09-05 12:45:00	2019-09-05 12:49:00	2019-09-05 10:00:00	2019-09-05 10:10:00	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮

1.2.3　特征编码

数据集中含有较多object类型的特征，为便于后续的运算与建模，需先对其进行特征编码。由于部分特征含有的类别数量较多，例如“飞机尾翼号”的类别有4 139个，如果采用one-hot encoding对进行编码，特征空间会变得过大，容易造成维度灾难。本文选择label encoding进行编码，将特征均转化为数值型。部分特征类型如表3所示。

表3　特征类型（部分）

Table 3　Type of feature (Part)

特征	类型
Carrier	Int32
WindDirection	Int32
DestID	Int64
FlightNumber	Int64
Distance	Float64
RelativeHumidity	Float64
⋮	⋮

1.3　特征选择

数据集中的冗余特征与无关特征会增加模型的计算量，减慢训练速度，甚至有产生过拟合的可能。对这些特征进行筛选，可以减少不必要的资源消耗，提升模型的预测性能。本文的特征选择主要分为两个部分：方差过滤与递归特征消除。

方差过滤是对所有特征的方差进行计算，并根据设定的阈值来过滤掉那些方差较小的特征。如果一个特征本身的方差很小，就代表着样本在这个特征上的大多数取值基本没有差异，甚至完全相同。例如在“出发机场名称”这一特征中，由于本文只选择了一个目标机场，故样本的取值也只有一种，这对于样本的区分毫无帮助。

递归特征消除是通过选定的基模型来对特征的重要性进行排序，在每一轮训练过程中，都消除掉一个或一些权重较小的特征，如此迭代进行，直至最后留下的特征个数满足要求。本文选用LightGBM作为基模型，在方差过滤的基础上进行递归特征消除，最终选定30个特征，其中航班信息相关特征21个，天气信息相关特征9个，部分特征选择结果的描述如表4所示。

表4　特征选择结果（部分）

Table 4　Feature selection results (Part)

特征名称	特征描述	方差值
Month	月份	8.52
Carrier	航空公司	4.75
DestID	目的机场	2 472 982.52
FlightNum	航班号	2 677 607.09
TailNum	飞机尾翼号	1 186 194.24
CrsElapsedTime	计划飞行时间	8 485.19
CrsDepTime	计划起飞时间	194 697.96
PriorArrDelay	前序航班延误时间	8 407.36
SkyConditions	云层状况	3 016 787.20
Visibility	能见度	4.501 1
WindSpeed	风速	24.28
WeatherType	异常天气类型	249.81
⋮	⋮	⋮

1.4 不平衡处理

本文将航班延误的严重程度分为5级[8]，按照航班的离港延误时长进行划分，具体方式如表5所示，其中t为实际起飞时间与计划起飞时间之差。

表5 航班延误等级划分

Table 5 Classification of flight delay

延误等级	延误时间/min	分类编码
未延误	$t\leqslant 15$	0
轻度延误	$15<t\leqslant 60$	1
中度延误	$60<t\leqslant 120$	2
高度延误	$120<t\leqslant 240$	3
重度延误	$t>240$	4

每类样本所占总样本的比例如图1所示。从图1中可以看出，未延误航班的数量接近总航班数量的3/4，约为占比最小的重度延误航班的75倍。如果直接对这样的数据集进行预测，很可能会造成多数类样本过拟合，而其他类样本欠拟合的结果，模型也会更偏向于将样本预测成为“未延误”航班，无法达到对航班的延误等级进行精准预测的效果。

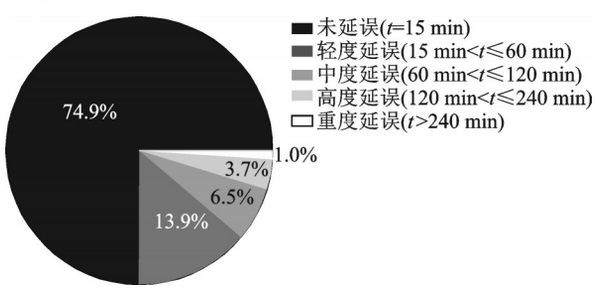

图1 不平衡处理前各延误等级航班占比

Fig.1 Proportion of flights with different delay levels before imbalance treatment

对不平衡数据进行处理，主要是通过重采样的方法调整原始数据中每个类别的样本数量，使各类别的样本数相对均衡。本文模型采用的SMOTE-Tomek组合采样，即先使用合成少数过采样技术(Synthetic minority oversampling technique，SMOTE)算法，通过少数类样本的最近邻来随机生成新样本，再移除数据中的Tomek link，在各类样本大致均衡的前提下，尽量保持分类边界的清晰。处理过后的数据各等级分布比例如图2所示。

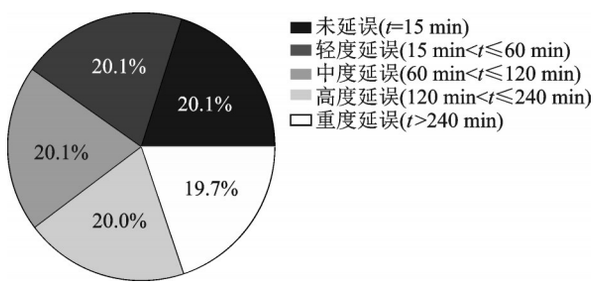

图2 不平衡处理后各延误等级航班占比

Fig.2 Proportion of flights with different delay levels after imbalance treatment

2 基于LightGBM的航班延误预测模型

2.1 LightGBM算法介绍

2.1.1 算法原理

LightGBM是梯度提升决策树(Gradient boosting decision tree，GBDT)的一种高效实现。它的原理与GBDT相似，是将损失函数的负梯度作为当前决策树的残差近似值，去拟合新的决策树，即每一次迭代都保留原来的模型不变，再加入一个新的函数到模型中，使预测值不断逼近真实值。

训练的目标函数如式(1)所示，其中，y_i为标签的真实值，$\hat{y}_i^{K-1}$为第$K-1$次学习的结果，c^{K-1}为前$K-1$棵树的正则化项和，目标函数的含义为寻找一棵合适的树f_k使得函数的值最小。

$$\mathrm{Obj}^K=\sum_i L(y_i,\hat{y}_i^K)+\Omega(f_k)+c^{K-1}=\sum_i L(y_i,\hat{y}_i^{K-1}+f_K(x_i))+\Omega(f_k)+c^{K-1} \tag{1}$$

运用泰勒公式对目标函数进行展开

$$f(x+\Delta x)=f(x)+f'(x)\Delta x+\frac{1}{2}f''(x)\Delta x^2 \tag{2}$$

损失函数的二阶泰勒展开结果为

$$\sum_i L(y_i,\hat{y}_i^{K-1}+f_K(x_i))=\sum_i[L(y_i,\hat{y}_i^{K-1})+L'(y_i,\hat{y}_i^{K-1})f_K(x_i)+\frac{1}{2}L''(y_i,\hat{y}_i^{K-1})f_K^2(x_i)] \tag{3}$$

用g_i记为第i个样本损失函数的一阶导数，h_i记为第i个样本损失函数的二阶导数

$$g_i=L'(y_i,\hat{y}_i^{K-1}) \tag{4}$$

$$h_i=L''(y_i,\hat{y}_i^{K-1}) \tag{5}$$

简化后的目标函数可表示为

$$\mathrm{Obj}^K=\sum_i[L(y_i,\hat{y}_i^{K-1})+g_i f_K(x_i)+\frac{1}{2}h_i f_K^2(x_i)]+\Omega(f_K)+c \tag{6}$$

2.1.2 算法优势

传统的GBDT算法在构建决策树时，选用的是Pre-sorted算法来寻找最优分割点，对每个特征都要遍历其所有的数据样本，计算所有可能分割点的信息增益。如图3所示，LightGBM采用了改进的Histogram算法，将连续的浮点特征值划分为k个区间，只需要在这k个区间中选择最优分割点，大大提升了训练速度与空间的利用效率[7]。

除此之外，LightGBM从减少训练数据的角度，在建立决策树时采用按叶生长(Leaf-wise)策

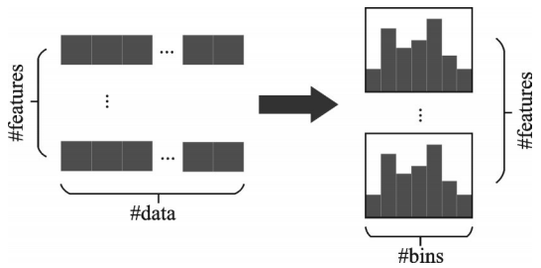

图3　直方图算法原理图

Fig.3　Schematic diagram of histogram algorithm

略代替按层生长(Level-wise)策略(图4)，并增加最大深度的限制，在保证高效率的同时防止过拟合。采用单边梯度采样(Gradient-based one-side sampling，GOSS)保留梯度较大的实例，对梯度较小的实例进行随机抽样，用更小的数据量获得精确的信息增益估计。从减少特征的角度，采用互斥特征合并(Exclusive feature bundling，EFB)将一定的冲突比率内互斥的特征进行合并，达到降维的效果，且不会造成信息丢失[7]。

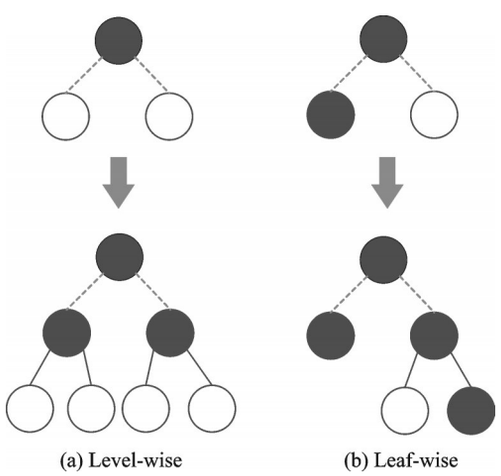

图4　Level-wise策略与Leaf-wise策略

Fig.4　Level-wise strategy and Leaf-wise strategy

2.2　预测流程

本文提出一种基于LightGBM的航班延误多分类预测模型，以纽瓦克自由国际机场为目标机场，收集相关的航班数据与天气数据，并按照时间标签进行融合。对数据进行预处理，主要包括缺失值处理、前序航班信息处理和特征编码等，并运用方差过滤以及递归特征消除进行特征选择，最终的数据集共包含30个特征。预测的标签按照延误时长进行划分，共5个等级，采用SMOTE与Tomek Link对数据进行重采样，改善其不均衡特性。最后划分训练集与测试集，使用LightGBM算法进行多分类预测，经贝叶斯调参得出最终模型，并用测试集进行预测，根据结果来对模型性能进行评估。算法预测流程如图5所示。

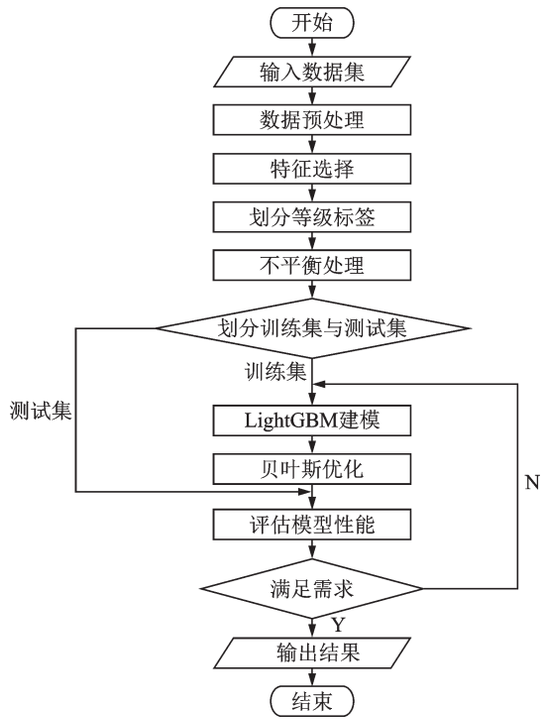

图5　基于LightGBM的航班延误多分类预测流程

Fig.5　Multi-classification forecast process of flight delay based on LightGBM

2.3　贝叶斯调参

本模型调参方式选用贝叶斯优化，即寻找可以使目标函数达到全局最大的参数时，会考虑已有的先验信息，从而更好地调整当前的参数[4]。相比于网格调参，贝叶斯参数迭代次数少，运行速度更快，可以一次调整多个参数，不容易造成维度爆炸，且只需要给参数制定大体的调整范围，不需要考虑如何对范围进行进一步细分。

数据集共包含样本1 156 413条，其中75%的数据作为训练集，余下25%的数据作为测试集。设定调参范围，并同时对最大叶子数、最大深度、学习率、最小分裂增益样本抽样率和特征抽样率等多个参数进行调参，最终结果如表6所示。

表6　贝叶斯优化的调参范围与结果

Table 6　Parameter adjustment range and results of Bayesian optimization

参数名称	调参范围	调参结果
Num_leaves	(150,300)	273
Max_depth	(12,16)	14
Learning_rate	(0.6,1.5)	0.6
Min_Split_gain	(0,1)	0.039 0
Bagging_fraction	(0.6,1)	0.692 7
Feature_Fraction	(0.6,1)	0.604 2
Max_bin	(100,256)	191

3 实验结果与分析

3.1 评价指标

用于评价分类模型的性能指标主要包括以下4种:准确率、精确率、召回率以及F_1分数。其中准确率是预测正确的结果占总样本的百分比,代表对样本整体的预测准确程度。精确率是被所有预测为正的样本中实际为正样本的概率,代表正样本结果中的预测准确程度。准确率与精确率的指标定义如下,其中TP为正样本被判断为正,TN为负样本判断为负,FP为负样本判断为正,FN为正样本判断为负。

$$A=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{7}$$

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{8}$$

式中:A为准确率;P为精确率。

在样本不均衡的情况下,不能只参照准确率对模型进行评估。在正样本的数量远远少于负样本时,即使将所有样本都预测为负类样本,也可以达到很高的准确率,但模型并没有起到任何检测正样本的作用。因此,在此类问题中,需要同时参考召回率这一指标。召回率是在实际为正的样本中被预测为正样本的概率,代表着对少数类样本的捕捉能力,在航班延误问题中,也就代表着对少数延误航班的检测能力。

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{9}$$

式中R为召回率。

F_1分数可以理解为精确率与召回率的调和平均数,综合了精确率和召回率的结果,能够客观全面地反映模型性能,其数值最小为0,越接近1代表模型的性能越好。

$$\frac{2}{F_1}=\frac{1}{P}+\frac{1}{R} \tag{10}$$

如上所述的对于精确率、召回率与F_1分数的传统计算公式只适用于二分类模型,由于本文所采用的模型为多分类,所以按照Macro average规则来进行计算,即分别计算每个类别精确率、召回率与F_1,然后求均值,平等地对待每个类别。用于多分类的准确率、精确率与F_1分数的指标定义如下

$$P_{\mathrm{Macro}}=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{11}$$

$$R_{\mathrm{Macro}}=\frac{1}{n}\sum_{i=1}^{n}R_i \tag{12}$$

$$F_{\mathrm{Macro}}=\frac{1}{n}\sum_{i=1}^{n}F_i \tag{13}$$

$$F_{\mathrm{Macro}}=\frac{2\times P_{\mathrm{Macro}}\times R_{\mathrm{Macro}}}{P_{\mathrm{Macro}}+R_{\mathrm{Macro}}} \tag{14}$$

3.2 预测结果

本文将BTS所提供的航班数据与NOAA所提供的天气数据相结合,经过上述处理,形成一份包含1 156 413个样本,30个特征变量的数据集。类别标签以航班延误的时长为标准,划分为从0~4的5个延误等级。选取数据的75%作为训练集,25%作为测试集,对航班延误进行多分类预测。

模型的迭代次数为120次,运行时长为2 min 16 s。最终预测结果的准确率为90.33%,精确率为90.30%,召回率为90.31%,F_1分数为0.902 4。

对预测结果的混淆矩阵进行可视化,得到的结果如图6所示。混淆矩阵是机器学习中用来总结分类模型预测结果的一种分析表,表中的每列代表预测类别,每行代表数据的真实类别,对角线的数值则代表被预测正确的样本数量。混淆矩阵对角线的数值越大,即混淆矩阵图对角线的颜色越深,模型的分类性能越好。

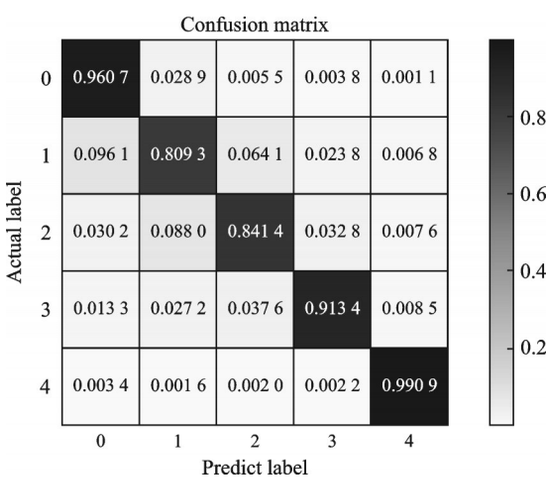

图6 混淆矩阵图

Fig.6 Confusion matrix graph

提取在预测过程中,重要程度在前15名以内的特征,如图7所示。特征分别为云层状况、机尾号、航班号、计划起飞时间、计划飞行时间、前序航

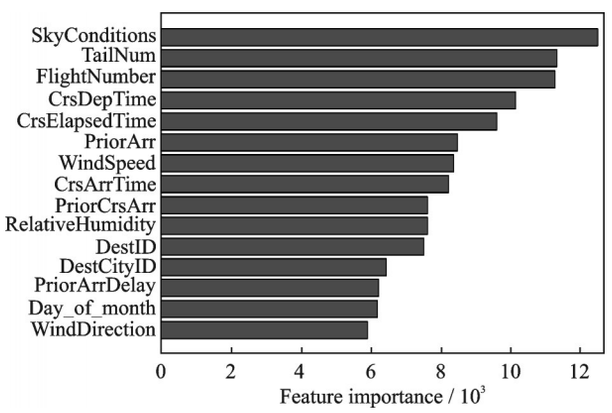

图7 不同特征的重要性分布

Fig.7 Importance distribution of different features

班实际到达时间、风速、前序航班预计到达时间、相对湿度、目的机场、目的城市、前序航班延误时间、每月第几天、风向，其中天气特征约占总重要特征的26.7%。

3.3　分类实验

为验证处理步骤的合理性，构建多个数据集，并对实验结果进行对比。实验1为3.2节所训练的模型，即使用经过不平衡处理、前序航班特征处理且含有天气信息的数据集，实验2为仅缺失天气信息的数据，实验3为仅缺失前序航班特征的数据，实验4为仅缺失不平衡处理的数据。模型的参数与迭代次数均相同，在测试集上的性能表现如表7所示。

表7　实验结果对比

Table 7　Comparison of experimental results

实验	准确率/%	精确率/%	召回率/%	F_1分数
1	90.33	90.30	90.31	0.902 4
2	83.89	83.73	83.89	0.837 1
3	83.47	83.21	83.45	0.830 7
4	82.47	69.05	47.98	0.548 6

实验结果证明，在分别增加了前序航班特征与天气数据后，航班延误预测模型的各项性能都提升了6%以上，充分说明了前序航班特征与天气数据对航班延误预测模型的提升起到了良好作用。未经过不平衡处理的数据集，虽然准确率基本达到了80%以上的水平，精确率、召回率与F_1分数却大幅度降低，尤其是召回率仅为47.98%，意味着模型对延误航班的检测能力相当有限，这一实验结果可以更直观地通过混淆矩阵图来比较。

对比图6与图8可以看出，在未经过不平衡处理时，混淆矩阵图的第1列颜色很深，意味着模型更偏向于将更高等级的延误航班预测为“未延误”，即样本数量占比较多的多数类。在经过不平衡处理后，第1列的颜色趋于正常，对角线的颜色加深，意味着预测正确的样本数量增多，模型对延误航班的分类性能变好。对比结果证明了对数据进行不平衡处理的的必要性。

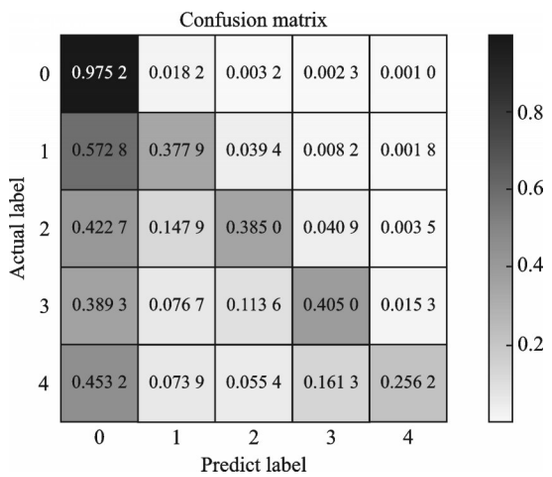

图8　不平衡处理前混淆矩阵图

Fig.8　Confusion matrix graph before unbalance treatment

3.4　不同算法对比

为进一步对本文算法所实现的航班延误多分类预测性能进行评估，将其与较为先进的XGBoost、GBDT与Ramdom Forest算法相比较。不同算法在训练时的最大深度均为15层，在测试集上的性能表现如表8所示。

在对相同数据集的处理中，本文算法在准确率、精确率、召回率以及F_1分数4大指标中，均是最优秀的，且在保持良好性能的同时，大幅度地降低了时间成本。本文算法对1 156 413条数据进行分析处理，并达到90%以上的准确率，仅需要花费2 min 16 s的时间，而XGBoost达到了82%以上的准确率，需要花费6 min 31 s的时间，是LightGBM的2.875倍，GBDT与Random Forest所需时间更久。实验结果证实了本文算法在航班延误的多分类预测问题中，预测性能与训练速度均优于其他算法。

表8　不同算法的实验结果对比

Table 8　Comparison of experimental results of different algorithms

算法	准确率/%	精确率/%	召回率/%	F_1分数	时间
本文算法	90.33	90.30	90.31	0.902 4	2 min 16 s
XGBoost	82.65	82.43	82.63	0.822 9	6 min 31 s
GBDT	72.75	71.89	72.73	0.721 3	14 min 39 s
Ramdom Forest	67.66	67.23	67.65	0.667 0	7 min 30 s

4　结　　论

提前对航班延误的严重程度进行预测，有助于将事后被动应急转为事前主动干预，减缓延误累积的负面影响。本文根据真实的航班数据，提出了一种基于LightGBM的航班延误多分类预测模型。主要工作有：(1)将天气信息与航班信息相结合，并增加前序航班的相关特征，综合考虑各个因素对航班延误的影响；(2)运用方差过滤与递归特征消除，对无关特征与冗余特征进行筛选，降低模型复杂程度与运算成本；(3)综合运用SMOTE与Tomek Link对数据进行重采样处理，消除数据的不平衡特性；(4)通过LightGBM算法与贝叶斯优化对航班延误时长进行多分类预测，并对模型进行全面的评估与比较。实验结果表明，相比与其他先进算法所构建的预测模型，本文模型具有更低的训练时间成本与更精准的预测性能，可以为航班延误

的预测问题提供高效准确的参考。未来的研究工作将会考虑空域限制、流量管理以及到达机场天气等因素,进一步提升航班延误预测的准确率。

参考文献:

[1] KLEIN A, CRAUN C, LEE R S. Airport delay prediction using weather-impacted traffic index (WITI) model[C]//Proceedings of 2010 IEEE/AIAA 29th Digital Avionics Systems Conference. Salt Lake City, USA:IEEE, 2010: 1-13.

[2] RODRIGUEZ-SANZA A, GOMEZ COMENDADOR F, ARNALDO V, et al. Assessment of airport arrival congestion and delay: Prediction and reliability[J]. Transportation Research Part C Emerging Technologies, 2019, 98(1): 255-283.

[3] ÁLVARO R S , FERNANDO G C, ROSA A V, et al. Assessment of airport arrival con-gestion and delay: Prediction and reliability[J]. Transportation Research Part C: Emerging Technologies, 2019, 98(1): 255-283.

[4] WU W, CAI K, YAN Y, et al. An improved SVM model for flight delay prediction[C]//Proceedings of 2019 IEEE/AIAA 38th Digital Avionics Systems Conference. San Diego, USA:IEEE, 2019: 1-6.

[5] ACHENBACH A, SPINLER S. Prescriptive analytics in airline operations: Arrival time prediction and cost index optimization for short-haul flights[J]. Operations Research Perspectives, 2018, 5(1): 265-279.

[6] GUI G, LIU F, SUN J, et al. Flight delay prediction based on aviation big data and machine learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(1): 140-150.

[7] 周洁敏,戴美泽,卢朝阳,等. 基于弹性神经网络的航班延误时间预测[J]. 航空计算技术,2019,49(5): 12-16.
ZHOU Jiemin, DAI Meize, LU Chaoyang, et al. Flight delay prediction based on elastic neural network[J]. Aeronautical Computing Technology, 2019, 49(5): 12-16.

[8] 吴仁彪,赵婷,屈景怡. 基于深度SE-DenseNet的航班延误预测模型[J]. 电子与信息学报, 2019, 41(6): 1510-1517.
WU Renbiao, ZHAO Ting, QU Jingyi. Flight delay prediction model based on deep SE-DenseNet[J]. Acta Electronica Sinica, 2019, 41(6): 1510-1517.

[9] KE G, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[C]//Proceedings of Advances in Neural Information Processing Systems. Long Beach, USA: Neural Information Processing Systems Foundation, 2017: 3149-3157.

[10] LIU Y, YU Z, CHEN C, et al. Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net[J]. Analytical Biochemistry, 2020, 609(22): 113903-113912.

[11] WANG Y, WANG T. Application of improved lightgbm model in blood glucose prediction[J].Applied Sciences, 2020, 10(9): 3227-3242.

[12] TANG M, ZHAO Q, DING S X, et al. An improved LightGBM algorithm for online fault detection of wind turbine gearboxes[J].Energies, 2020, 13(4): 807-822.

[13] JU Y, SUN G, CHEN Q, et al. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting[J]. IEEE Access, 2019, 7(1): 28309-28318.

[14] 崔佳旭,杨博. 贝叶斯优化方法和应用综述[J]. 软件学报, 2018, 29(10): 3068-3090.
CUI Jiaxu, YANG Bo. Review of Bayesian optimization methods and applications [J]. Acta Sinica Sinica, 2018, 29(10): 3068-3090.

(编辑:陈珺)