浙 江 大 學

大 数 据 期 末 论 文

题目　风机叶片结冰故障预测模型及其实现方法

所在小组　第 7 组

指导老师　赵春晖

课程名称　大数据解析与应用导论

所在学院　控制科学与工程学院

完成时间　2018 年 11 月

目 录

风机叶片结冰故障预测模型及其实现方法

曾之宸　林润泽　徐博文

（浙江大学 控制科学与工程学院，浙江 杭州　310027）

摘　要：

风机结冰问题严重影响了风力发电的效率，因此，对于风机结冰情况的预测，有助于提前掌握风机结冰的概率，并进行相应的处理，从而提高运行效率。

本文旨在通过数据分布特点分析、数据预处理、类间数据均衡化、冗余变量删减、特征提取与重构等措施，对原始数据集进行处理。在此基础上，采用多种分类预测模型，对于风机结冰情况进行分类预测。关于分类器的选择，采用逐步递进的方式，从线性模型拓展至非线性分类模型，包括 Fisher 判别法、随机森林分类算法、卷积神经网络等模型。

考虑到风机结冰问题的实际物理背景，不妨将原始时间分为“白天-黑夜”两类，对两类时间数据分别进行模型的构建，最终实现多模型融合的预测策略。由于测试集不含有时间标签信息，需要根据原始数据训练出“白天-黑夜”的判别模型，进而采用白天、黑夜两套模型对于风机结冰进行分类预测。

最后，对于数据处理结果以及预测算法进行横向与纵向的综合对比与分析。经过上述模型构建过程，实现了较好的结冰预测效果，具有一定的应用价值。

关键词：风机结冰预测；重采样；物理特征；Fisher 判别；随机森林；卷积神经网络；多模型融合

1 问题背景与重述

（1）背景介绍

叶片结冰是风电领域的一个全球范围难题。低温环境所导致的叶片结冰、材料及结构性能改变、载荷改变的问题等，对风机的发电性能和安全运行造成较大威胁。在这样的情况下运行会增加叶片折断损坏的风险。实际应用中面临的挑战是很难对结冰的早期过程进行精确预测，以便能够尽早开启除冰系统。

（2）问题重述

SCADA 系统每天产生大量的数据，但是目前大部分的系统依然局限于对已发生故障的报警。这些故障到达报警阶段时往往已经比较严重，需要对风机进行停机和维修，造成巨大的发电损失和维护成本。通过对 SCADA 系统产生的大数据环境进行挖掘和建模，能够对一些严重故障进行预测和诊断，从而使过去应激型的维护方式转变为主动预测型的维护方式，能够有效地改善风电设备的使用率和运维成本。

比赛数据集为多样本低维度的类不平衡数据集，且数据集中包含大量无效数据。显然，该问题是一个典型的二分类问题，考虑到风机结冰的时序相关性，需要引入连续时间序列下的分类方法。通过对于 28 维原始数据进行数据预处理，利用几种不同的分类预测算法和模型，最终获得风机结冰状况的预测结果，并进行数据预处理、特征工程、模型求解等方法的综合对比与分析。

为了使得问题的分析与求解更具逻辑性和严密性，不妨将问题分解为下列求解步骤：

问题 1：数据标签匹配及无效数据处理

问题 2：离群样本筛查与坏点剔除

问题 3：类间不均衡数据的均衡化策略选择与实现

问题 4：基于相关分析的冗余变量删减

问题 5：基于物理模型的特征提取与构造

问题 6：线性分类算法——Fisher 判别

问题 7：非线性分类算法——随机森林

问题 8：时序预测分类算法——卷积神经网络

问题 9：风机结冰预测算法的分析与对比

注意：问题求解过程中，我们采用 15 号风机数据作为训练数据、21 号风机数据作为测试数据，以验证预测模型的泛化能力。

2 数据分析与预处理

2.1 数据集介绍

本问题的原始数据来源于《第一届中国工业大数据竞赛》公开数据集，数据总长度为 2 个月，样本容量约 30 万，其中结冰/非结冰样本比例约为 1:10，是典型的类不平衡问题。根据相关特征展示如下：

表 2.1　原始数据集不同类数据的统计结果

数据集信息	15 号风机	21 号风机
样本总数	393886	190494
结冰数据	23846	10638
正常数据	350255	168930
无效数据	19785	10926

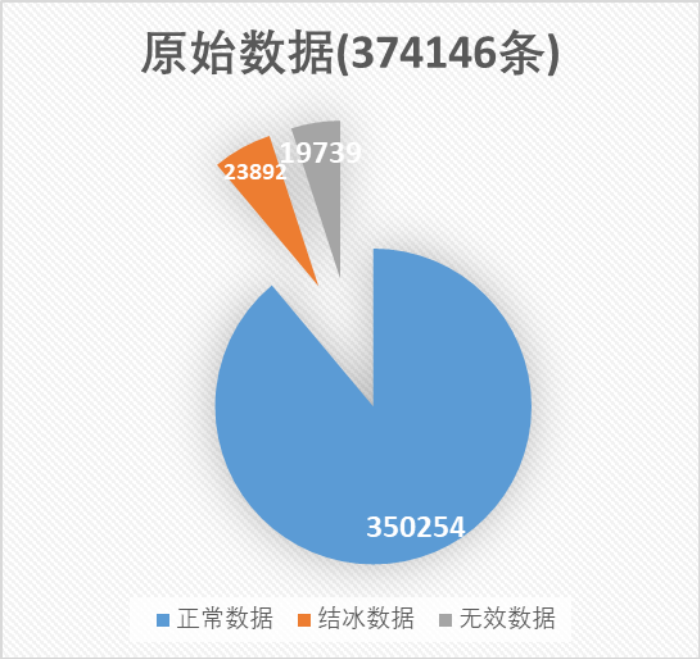


图 2.1　原始数据集 3 类数据量统计饼图

本问题的数据共包括 28 个连续数值型变量，涵盖了风机的工况参数、环境参数和状态参数等多个维度。变量的名称及说明，如表 2.1 所示。

表 2.1　原始数据集变量名称及其物理意义说明

字段名	说明
Time	时间戳
wind_speed	风速
generator_speed	发电机转速

Power	网侧有功功率(kw)
wind_direction	对风角(°)
wind_direction_mean	25 秒平均风向角
yaw_position	偏航位置
yaw_speed	偏航速度
pitch1_angle	叶片 1 角度
pitch2_angle	叶片 2 角度
pitch3_angle	叶片 3 角度
pitch1_speed	叶片 1 速度
pitch2_speed	叶片 2 速度
pitch3_speed	叶片 3 速度
pitch1_moto_tmp	变桨电机 1 温度
pitch2_moto_tmp	变桨电机 2 温度
pitch3_moto_tmp	变桨电机 3 温度
acc_x	x 方向加速度
acc_y	y 方向加速度
environment_tmp	环境温度
int_tmp	机舱温度
pitch1_ng5_tmp	ng5 1 温度
pitch2_ng5_tmp	ng5 2 温度
pitch3_ng5_tmp	ng5 3 温度
pitch1_ng5_DC	ng5 1 充电器直流电流
pitch2_ng5_DC	ng5 2 充电器直流电流
pitch3_ng5_DC	ng5 3 充电器直流电流
Group	数据分组标识

2.2 数据分布特点分析

2.2.1 存在离群坏点

分别对 26 个变量绘制“变量-时间”图，从图中可以观察到存在离群点的变量为：发电机转速（generator_speed）、偏航速度（yaw_speed）、ng5_2 温度（pitch2_ng5_tmp）、ng5_3 温度（pitch3_ng5_tmp）。

2.2.2 类间数据极度不平衡

（1）训练数据 15 号风机的样本总数 393886 条，其中，结冰数据 23846 条，

正常数据 350255 条，无效数据 19785 条。

（2）测试数据 21 号风机的样本总数 190494 条，其中，结冰数据 10638 条，正常数据 168930 条，无效数据 10926 条。

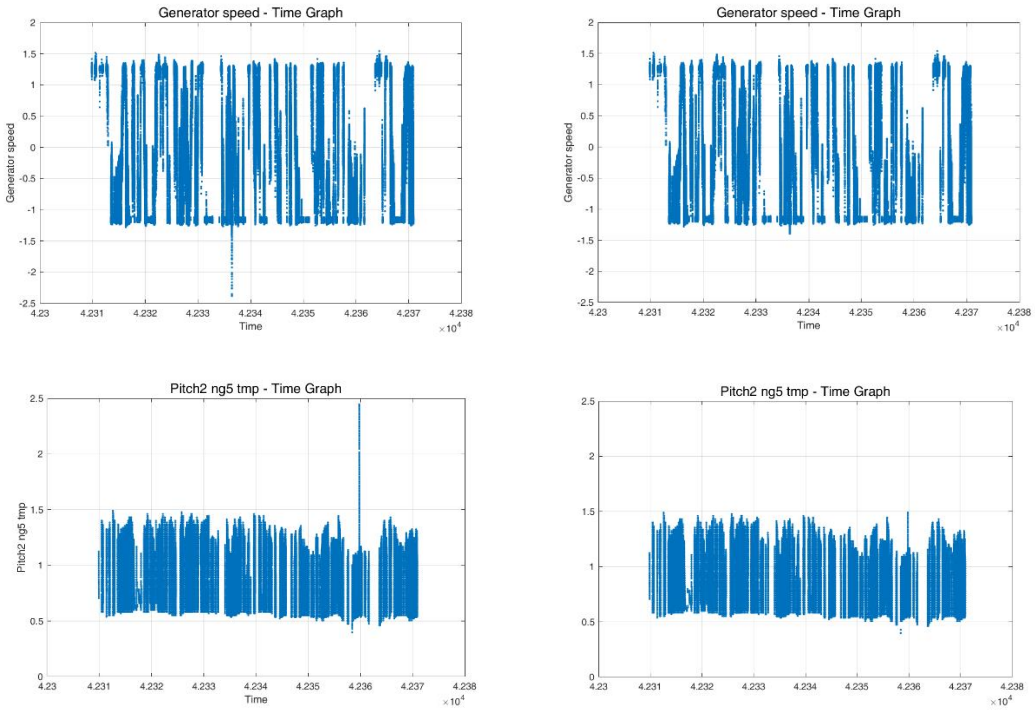
2.3 数据处理

2.3.1 数据标签匹配及无效数据处理

原始数据集包括三个 csv 文件（15_data.csv、15_failureInfo.csv 和 15_normalInfo.csv），其中，15_data.csv 包含 time 和其他 27 个特征，15_failureInfo.csv 为结冰工作时间段，15_normalInfo.csv 为正常工作时间段。15_data.csv 中的“time”维度代表的是每一条数据所对应的时间，需要依据 15_failureInfo.csv、15_normalInfo.csv 两个文件中的时间表，对相应的标签进行对比与匹配，以删除无效数据。

2.3.2 离群样本筛查与坏点剔除

根据之前绘制的“变量-时间”图，我们选择了合适的阈值对离群数据进行剔除，筛查结果如图 2.3.2 所示。



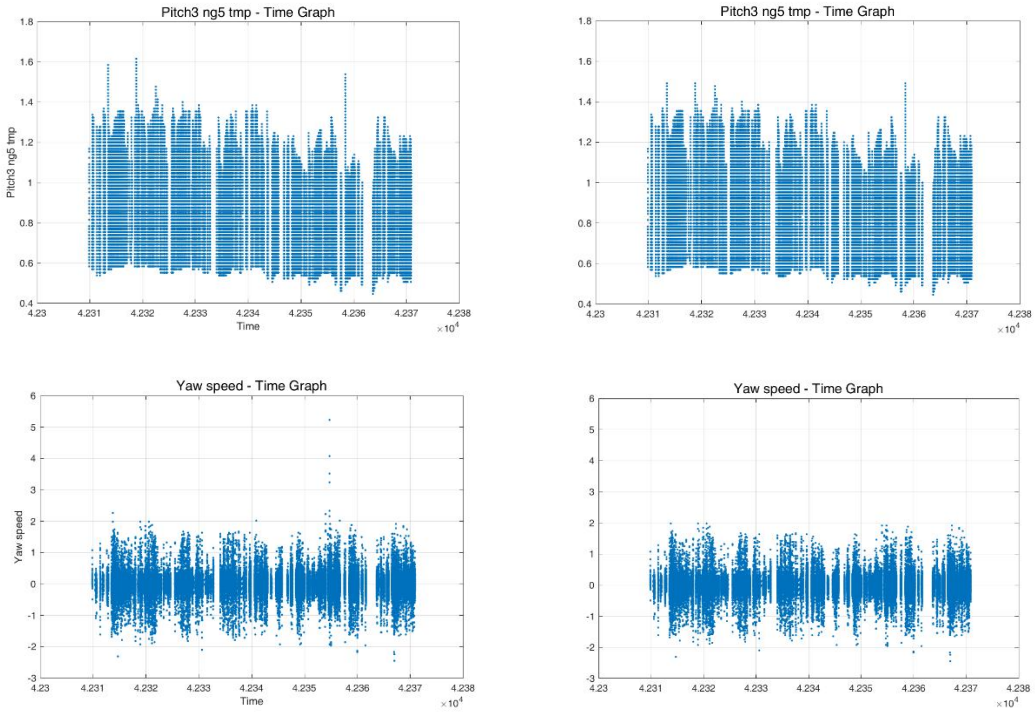


图 2.3.2　离群样本筛查与坏点剔除前后的数据-时间分布对比

（左边为原始数据分布情况，右边为离群点筛查后的数据分布情况）

2.3.3 类间不均衡数据的均衡化策略选择与实现

解决方案：

为消除类不平衡，一般采用大类样本降采样、小类样本过采样、同时改变两类样本权重、小类样本作为异常点等方法。权衡考虑到降采样带来的原始数据丢失，以及过采样引入的估计(非原始)数据，因此本文采取降采样过采样相结合的方法，尽可能规避单纯采用过采样或降采样带来的弊端。

其中过采样采用了 SMOTE 结合 ENN 的方式，降采样采用随机抽样（Random Under Sample）的删除方式。

SMOTE 算法基本思想是通过在少数类样本之间进行插值，从而获得额外的样本。具体地，对于一个少数类样本 X_i 使用 K 近邻法，求取距离 X_i 距离最近的 K 个少数类样本。本次求解中，我们采用样本之间 n 维特征空间的欧氏距离作为临近判据，从 K 个近邻点中随机选取一个，使用下列公式生成新样本：

$$X_{new} = X_i + (\hat{X}_i - X_i) \times \delta$$

其中，$\hat{X}_i$ 为选出的 K 近邻点，$\delta \in [0,1]$ 为一个随机数。

最近邻规则（Edited Nearest Neighbors，ENN）：ENN 算法基本思想是剔除离群的多数类样本。具体地，对于某一多数类样本，若其 K 个近邻点中有超过一半

属于少数类，则该样本会被剔除。该方法的另一个变种为，对于某一多数类样本，若其 K 个近邻点都不属于多数类，则这个样本会被剔除。

SMOTE 算法易导致生成的少数类样本与周围的多数类样本产生重叠，从而难以进行分类。而数据清洗技术恰好可以处理掉重叠样本，所以我们考虑将二者结合起来形成一个 pipeline，先进行过采样操作，再进行数据清洗，本文采用 SMOTE + ENN 结合的方法以清除更多的重叠样本，以使得过采样的样本能够更好地贴近原始数据。

基本原理：

随机降采样（Random Under Sample）基本思想是从多数类样本中随机选取一些剔除掉。该方法的缺点是被剔除的样本可能包含着一些重要信息，导致学习出来的模型效果欠佳；优点也比较明显，实现简单，能使样本快速达到平衡状态。

编程实现：

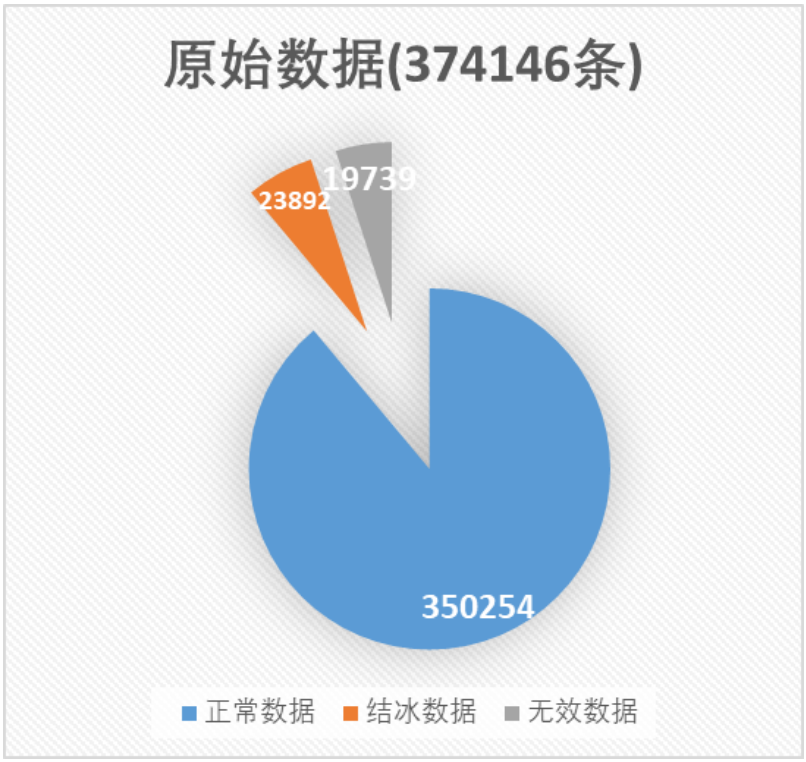
```
1. import numpy as np
2. import pandas as pd
3. import seaborn as sns
4. from imblearn.combine import SMOTEENN
5. from imblearn.under_sampling import RandomUnderSampler
6. X = pd.read_csv('test_data.csv')
7. y = pd.read_csv('test_labels.csv')
8. print("Reading Process Complete!")
9. smote_enn = SMOTEENN(sampling_strategy=0.4286,random_state=0)
10. X_resampled_1, y_resampled_1 = smote_enn.fit_resample(X, y)     #0.4
    286:1 过采样
11. print("Smote_enn Process Complete!")
12. randomundersampler = RandomUnderSampler(random_state=0)         #1: 1
    降采样
13. X_resampled_2, y_resampled_2 = randomundersampler.fit_resample(X_res
    ampled_1, y_resampled_1)
14. print("RandomUnderSampler Process Complete!")
15. X_resampled_smote = np.array(X_resampled_2)
16. y_resampled_smote = np.array(y_resampled_2)
17. np.savetxt('data_new.csv',X_resampled_smote,fmt='%e',delimiter=',')

18. np.savetxt('labels_new.csv',y_resampled_smote,fmt='%e',delimiter=','
    )
```

结果分析：

为了分析降采样和过采样对于分类预测模型的影响，不妨对原始数据集（记为数据集 1）、仅使用过采样得到的数据集（记为数据集 2）和使用过采样+降采样得到的数据集（记为数据集 3）。

首先，给出 3 个数据集的类间样本分布情况，如图 2.3.3 所示。



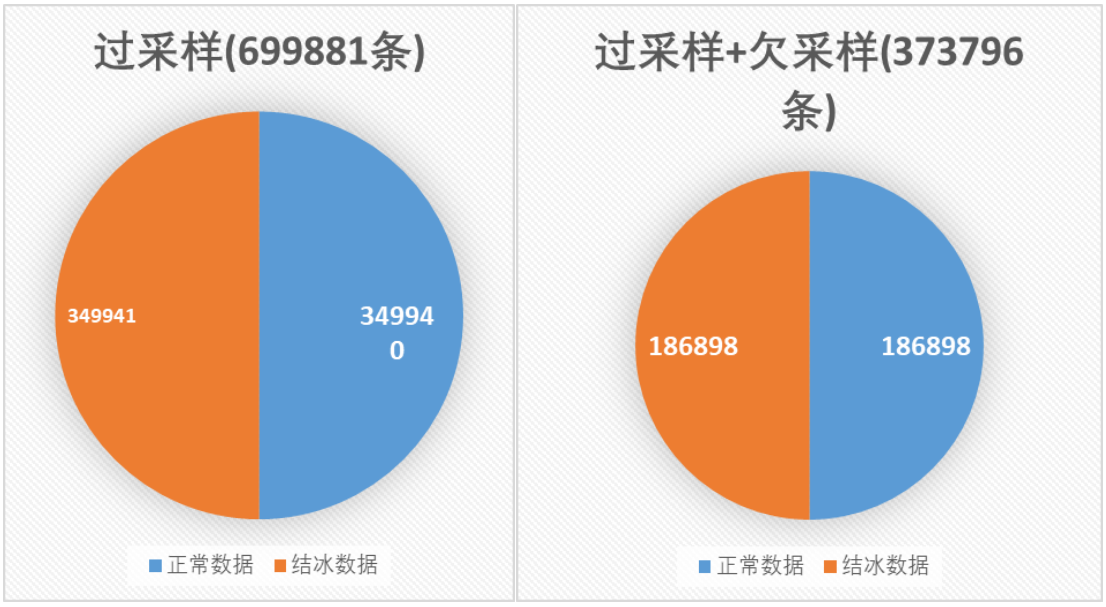
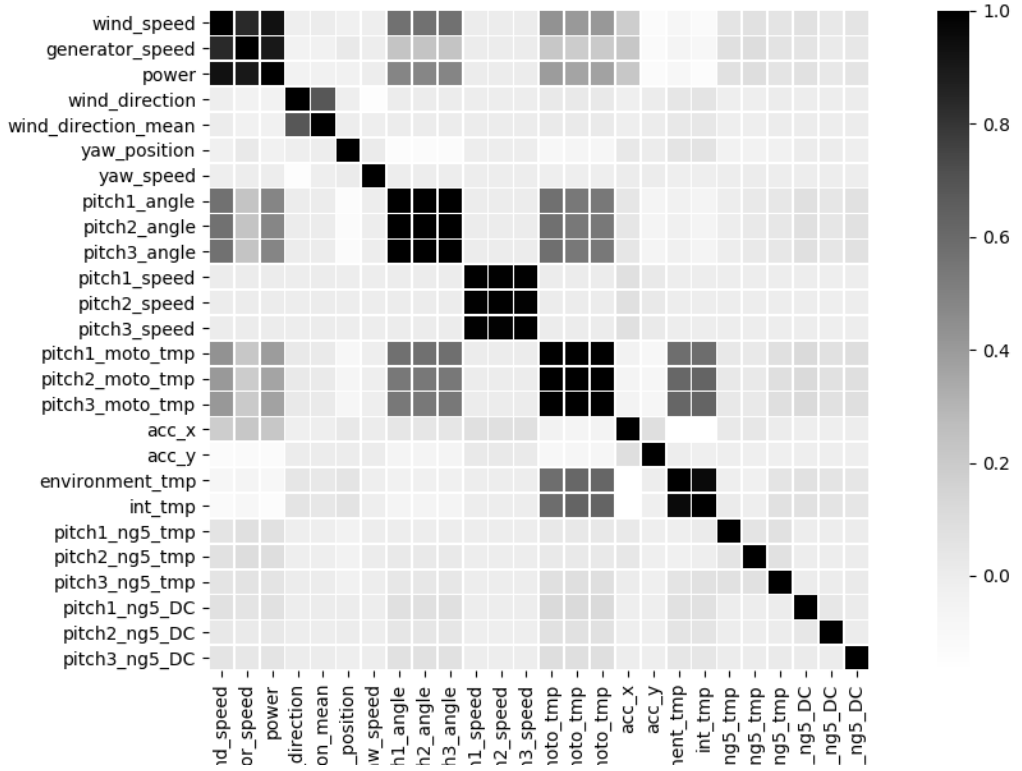
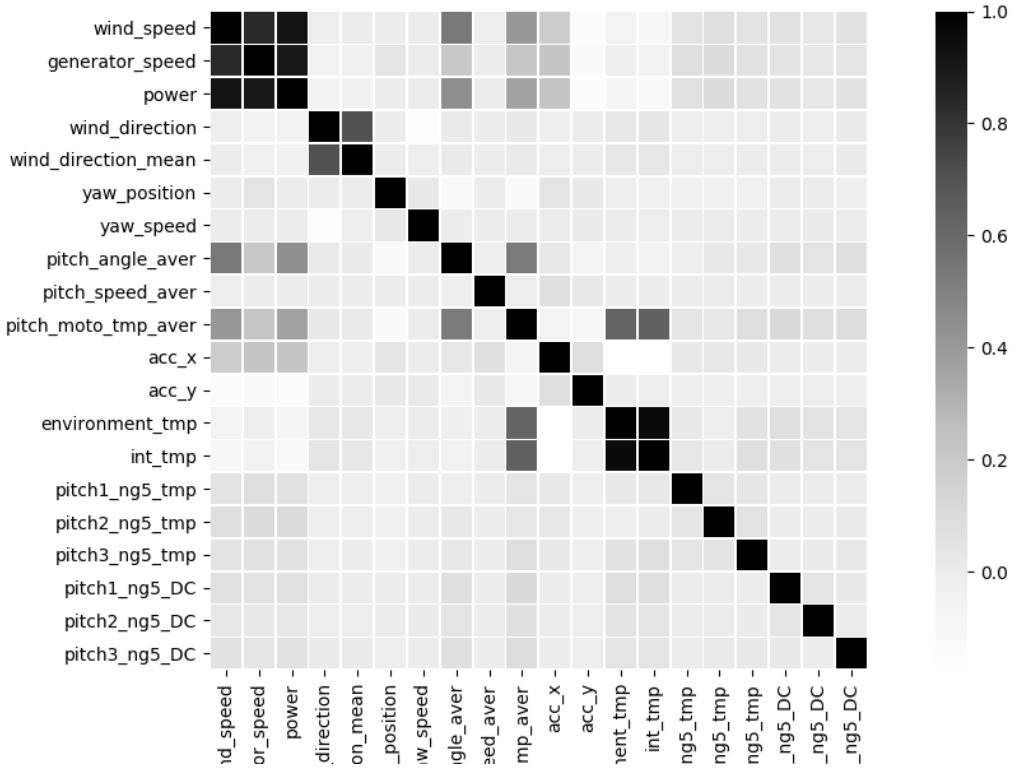


图 2.3.3.1 原始数据集与均衡化数据集类间样本数对比

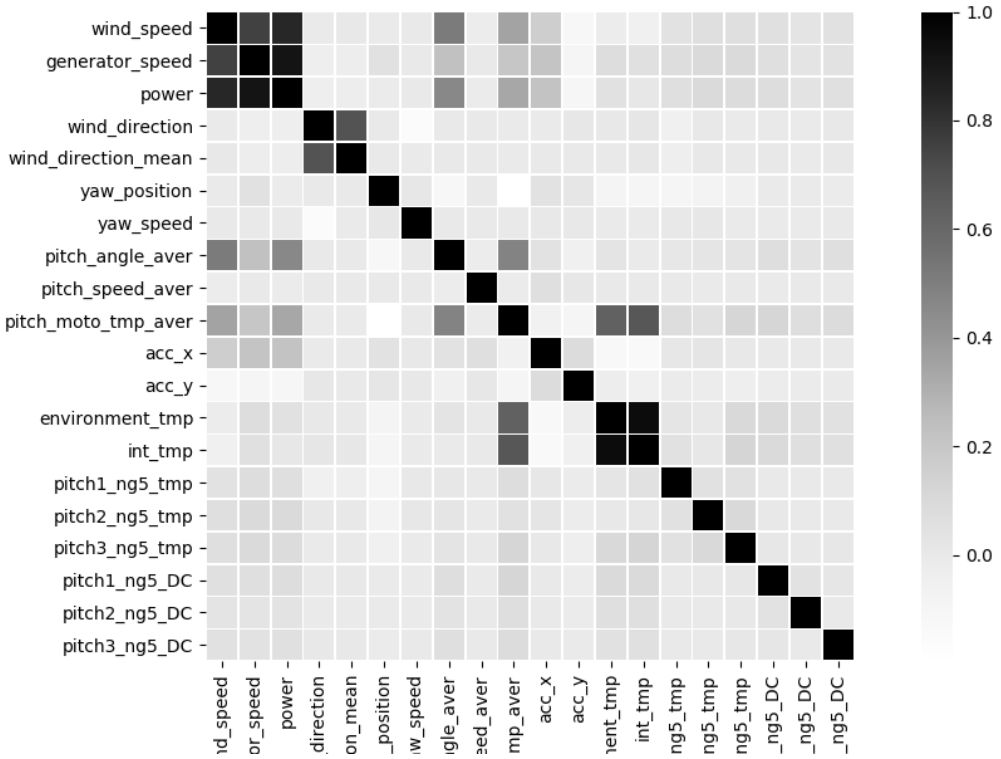
其次，为了分析类间均衡化对于数据集变量间耦合关系的影响，给出 3 个数据集的相关关系热力图，如图所示。



（a）数据集 1（原始）相关关系热力图



（b）数据集 2（过采样）相关关系热力图



（c）数据集 3（过采样+降采样）相相关关系热力图

图 2.3.3.2　原始数据集与均衡化数据集变量间相关关系热力图

通过相关关系热力图（图 2.3.3.2）可以看到，重采样之后变量相关关系基本无变化，我们认为采样操作并未使得采样后数据偏离原数据，所以采样后的结果是可信的。根据采样方式，我们将数据集分为三类：原始数据、仅通过过采样得获得的 80 万数据、过采样与降采样结合得到的 30 万数据。后续还将详细分析分别基于三类数据集进行分类预测结果的对比，此处不再赘述。

3 特征工程

3.1 基于相关分析的冗余变量删减

考虑到原始数据可能存在的耦合关系，在进行特征提取与特征重构之前，需要对数据预处理之后的数据集进行变量之间的相关关系分析，以便对于数据变量之前的牵连、耦合关系有一个直观的了解与认知。

基于上述分析，不妨绘制 26 个变量间的相关关系热力图，结果如图 3.1.1 所示。其中，相关关系热力图颜色的深浅表征了相关关系的强弱。

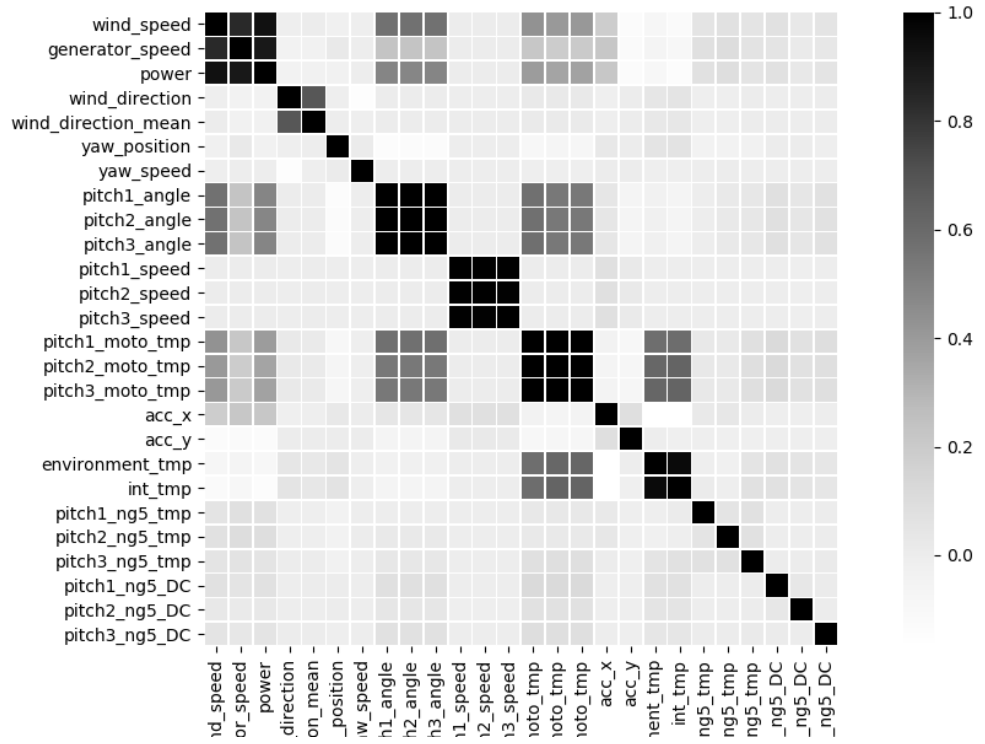


图 3.1.1 (a)　原始数据集变量相关关系热力图

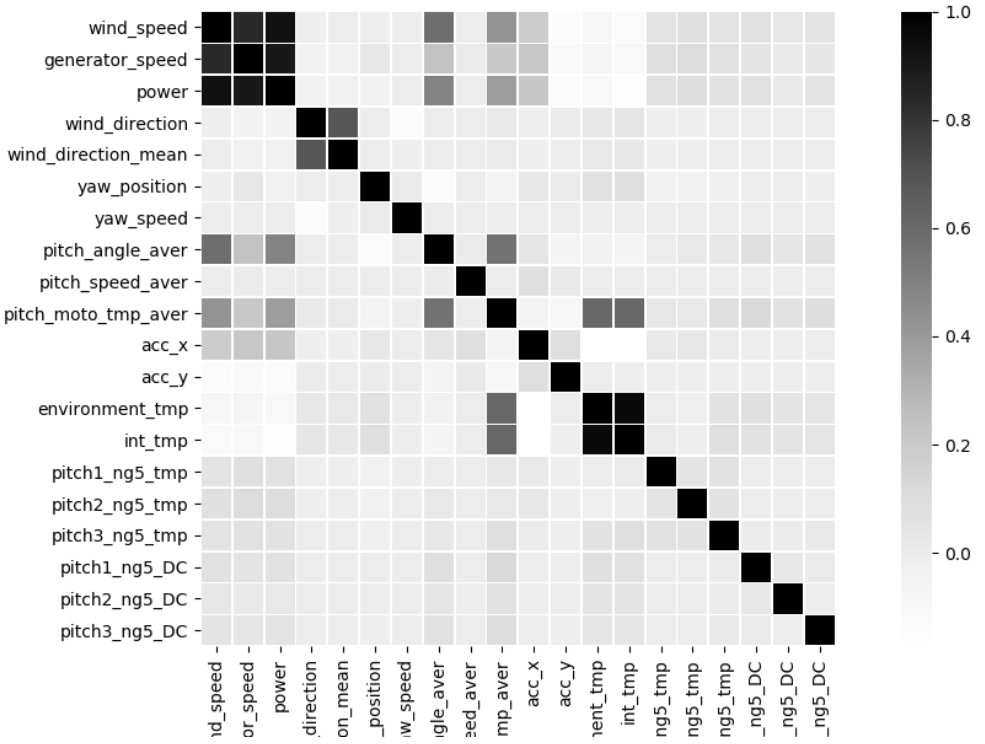


图 3.1.1 (b)　预处理后数据集变量相关关系热力图

根据相关关系热力图，可以看到，在 26 个变量中，其中的三组变量具有比较强的相关关系，分别为风机角度(1、2、3)、风机速度(1、2、3)、变桨电机温度(1、2、3)。除此之外，风速、发电机速度和网侧有功功率之间，以及环境温度、

机舱温度之间，也具有一定的相关关系。

为了更好地验证相关关系热力图所得结果的可靠性和正确性，需要进一步绘制上述三个变量两两之间的二维散点图，即风机角度(1、2、3)、风机速度(1、2、3)、变桨电机温度(1、2、3)变量之间的散点图，如图 3.1.2 所示。

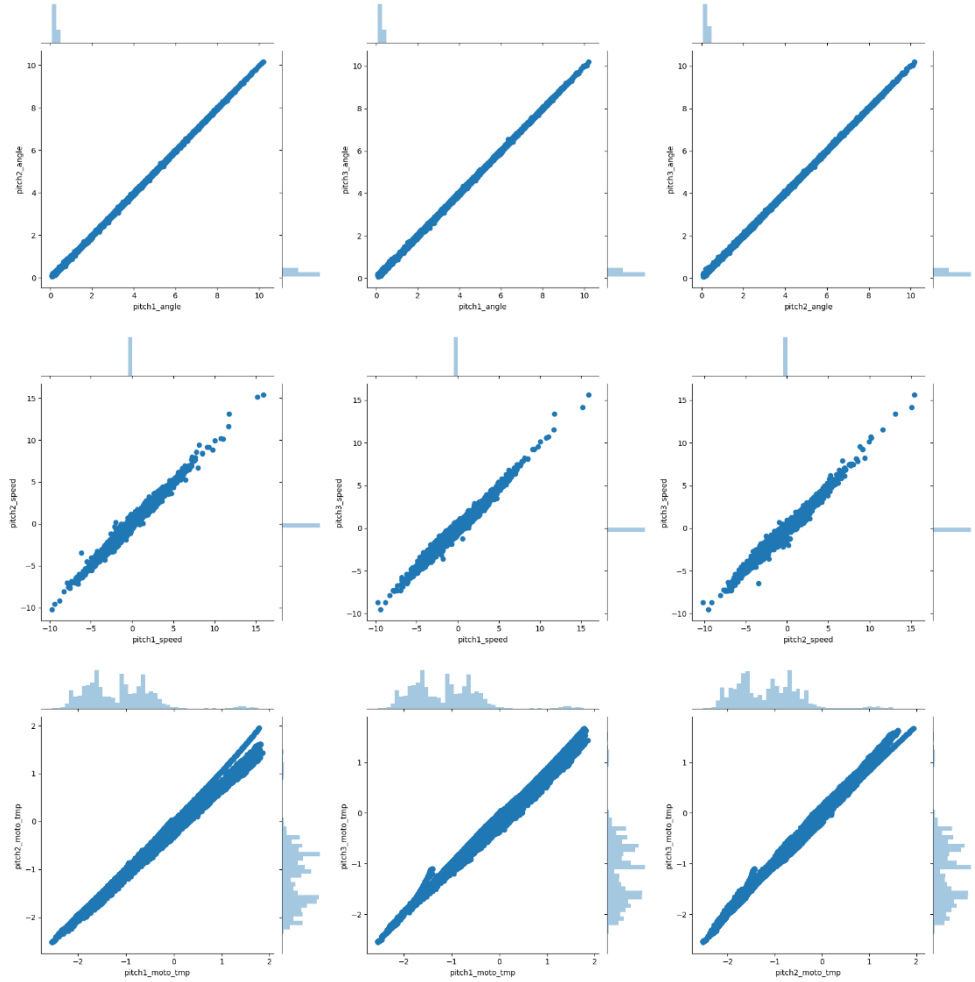


图 3.1.2　利用二维散点图验证相关关系分析的可靠性

从图 3.1.2 中，可以看出，三组变量两两之间的数据分布具有很强的线性关系，基本都围绕着主对角线在小范围内波动。值得一提的是，与风机三个叶片相关的三组对称参数具有非常高的对称性，这不仅体现在相关系数接近于 1，还体现在其分布规律具有惊人的相似性。

因此，根据相关分析和散点图绘制，并结合变量的物理含义，可以知道，同一风机 3 个叶片的相关性质具有一定的对称性，意味着三组变量为数据集中的冗余变量，完全可以用均值来替代原来的三个变量，以此作为新的特征，供后续的大数据建模方法求解的使用。

3.2 基于物理模型的特征提取与构造

由于风机结冰问题的分析和求解，纯碎依靠高性能算法来实现，具有很大的

局限性。因此，需要结合变量间的实际物理意义，对其中某些变量进行线性、非线性的组合运算，得到新的特征变量，用于大数据预测分类模型的构建。

根据相关文献的查阅，可以给出五个基于物理模型的特征变量的计算公式，如表 3.2 所示。

表 3.2　基于物理模型的特征提取与构造

特征变量	物理意义	计算公式
Tmp_diff	温差	$\text{int\_}tmp - environment\_tmp$
Torque	扭矩	$\frac{Power}{generator\_speed}$
Cp	功率系数	$\frac{Power}{wind\_speed^3}$
Ct	推力系数	$\frac{Torque}{wind\_speed^2}$
Lambda	速率比	$\frac{generator\_speed}{wind\_speed}$

问题求解过程中主要构造了温差、扭矩、功率系数、推力系数以及速率比五个物理特征，希望通过原始数据中的 26 个变量进行非线性组合，得到能够线性变量所无法表征的潜在关系。针对选取的物理变量，做说明如下：

- ✧ **温差：**表征环境温度与机舱温度的差值，温差绝对值越大，机舱结冰可能性越高；
- ✧ **扭矩：**表征风机转动所需克服的阻力，所需克服的阻力越大，扭矩越大，机舱结冰可能性越高；
- ✧ **功率系数：**表征风机发电功率与风速大小的相对关系，功率系数越低，机舱结冰可能性越高；
- ✧ **推力系数：**表征风力推动风机转动的阻尼程度，阻尼程度越高，推力系数越高，机舱结冰可能性越高；
- ✧ **速率比：**表征风机转速与风速的相对大小关系，速率比越低，机舱结冰可能性越高。

4 数据建模

4.1 分类器的选择依据

本课题研究过程中，我们采用递进的方式进行分类器的选择与优化。我们首先考虑使用以 Fisher 判别为代表的线性分类器，主要考虑变量之间的线性关系。根据相关关系热力图可以看到，特征工程之后的变量间线性相关关系较弱，所以单纯地使用线性分类器可能无法得到较好的分类结果。在认识到线性分类器在解决此问题上的乏力后，我们考虑使用以随机森林、卷积神经网络为代表的非线性

分类器。随机森林分类器在非线性问题的解决中有良好的表现，最终的分类结果没有让人失望，相比之前有了大幅的提升。随机森林算法并没有考虑数据的时序关系——这也是数据的重要特点。基于这个原因，我们选择了卷积神经网络分类器，获得了很好的效果。

4.2 线性分类算法——Fisher 判别

4.2.1 算法基本原理

Fisher 判别法是一种基于投影的线性分类方法，通过将高维空间的点向低维空间进行投影，从而实现降维操作。在原始数据坐标系下，可能难以将样品进行较好的分离，通过选择较好的投影方向，使得投影数据间类内紧缩、类间分离。一般说来，可以先将数据投影到一维空间（直线）上，当效果不理想时，再投影到另一条直线上，从而构成二维投影空间，依此类推，每个投影均可以建立一个相应的判别函数。

4.2.2 最终结果与对比

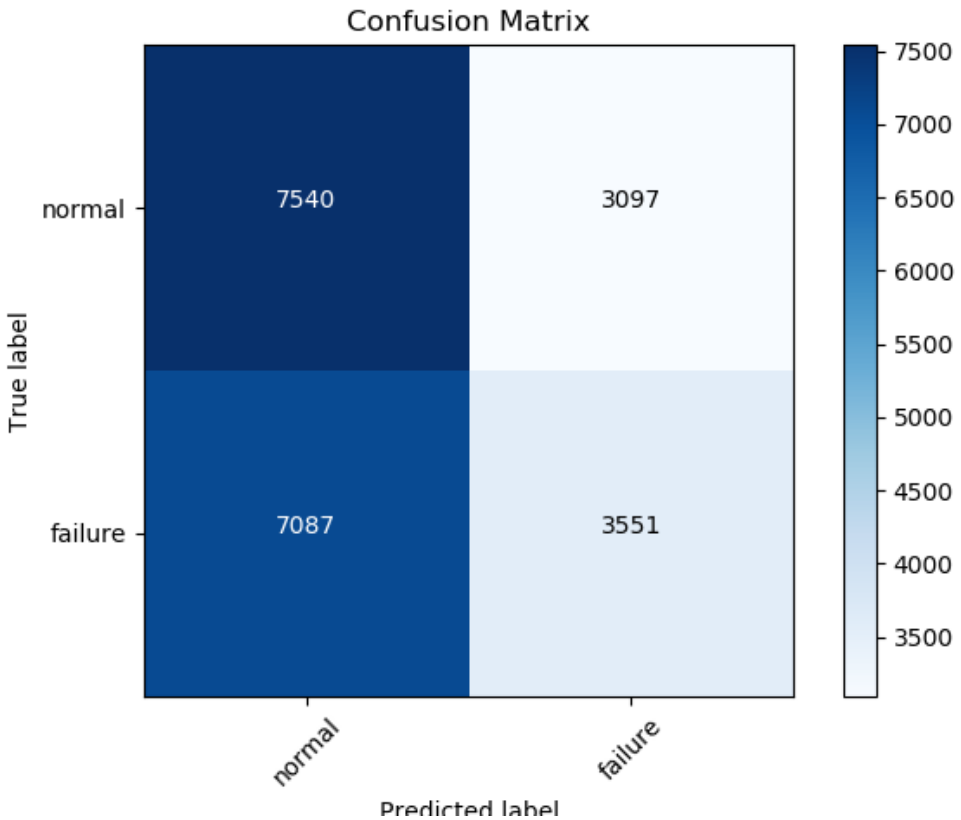


图 4.2.2(a)　基于原始数据集的 Fisher 判别算法结果

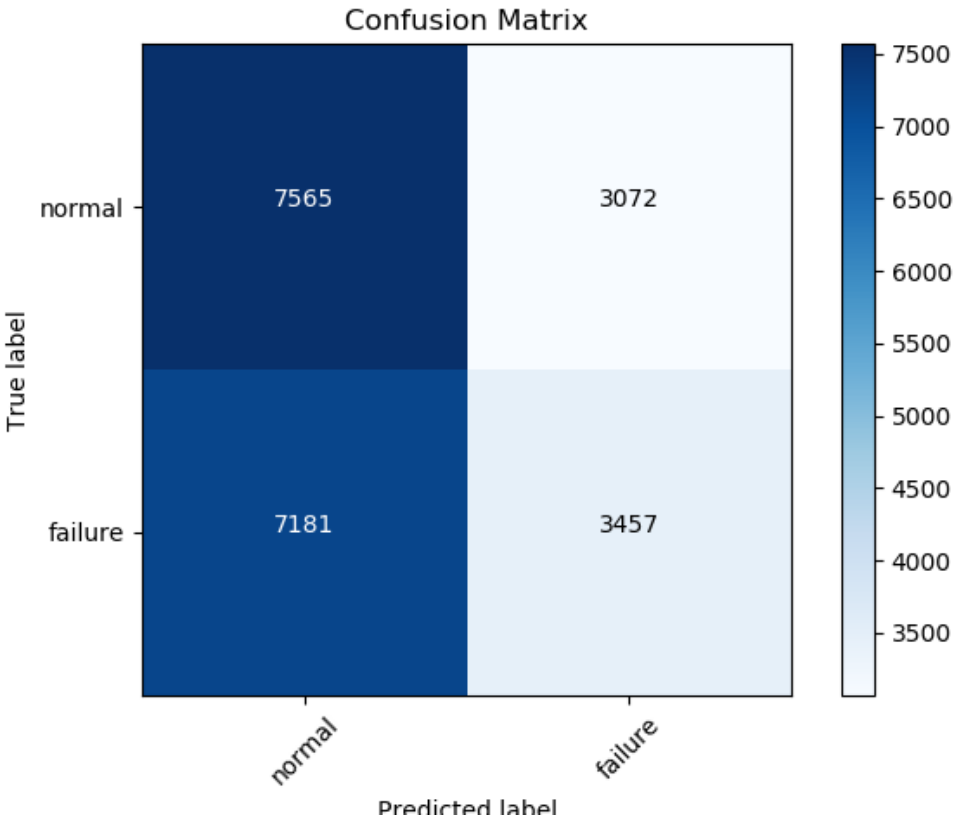


图 4.2.2(b)　基于过采样数据集的 Fisher 判别算法结果

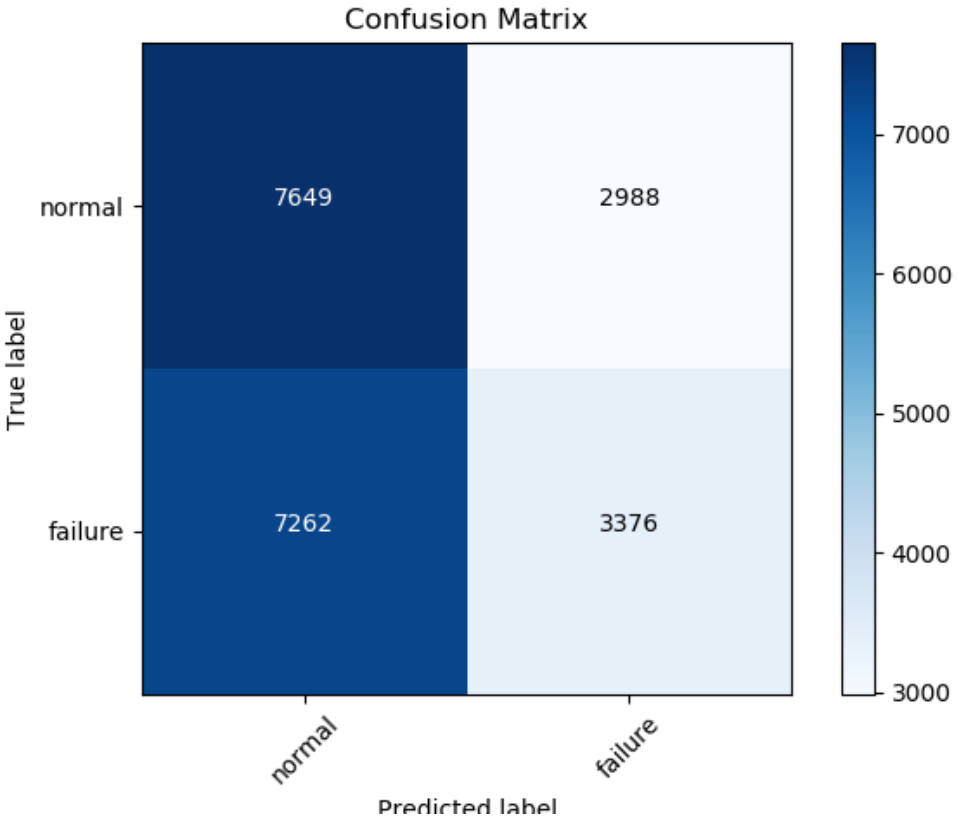


图 4.2.2(c)　基于过采样+降采样数据集的 Fisher 判别算法结果

4.3 非线性分类算法——随机森林

4.3.1 算法基本原理

决策树（Decision Tree）是在已知各种情况发生概率的情况下，通过构成决策树来求取净现值的期望值大于 0 的概率，是直观运用概率分析的一种图解法。决策树是一种带有特殊含义的树结构，其根结点（非叶子结点）代表数据的特征标签，根据该特征不同的特征值将数据划分成几个子集，每个子集均是这个根结点的子树，并对每个子树进行递归划分，而决策树的每个叶子结点则是数据的最终类别标签。对于一个样本特征向量，从决策树的顶端往下进行分类，直到根结点，得到的类别标签即为该样本向量的类别。

随机森林通过随机的方式建立一个森林，在森林中由很多决策树组成，并且每一棵决策树相互之间是没有关联的。每当有一个新样本，森林的每一棵决策树均会对其进行判断类别，通过投票的方式，选择得票最高的类别作为最终的分类结果。

4.3.2 最终结果与对比

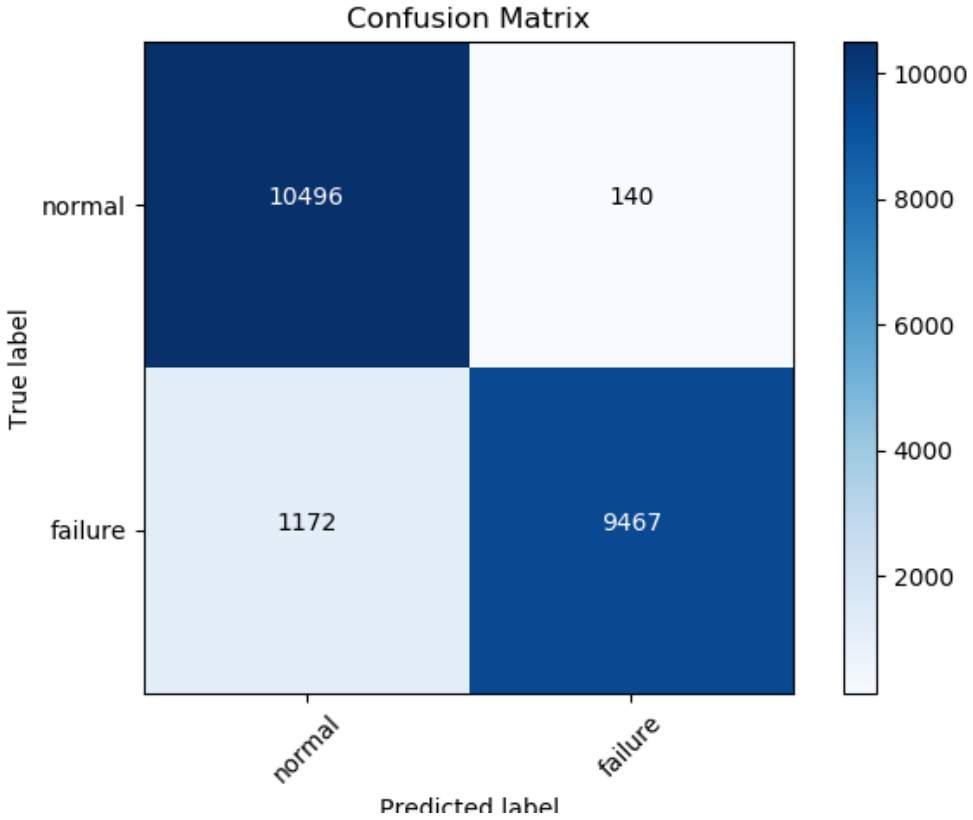


图 4.3.2(a) 基于原始数据集的 Random Forest 算法结果

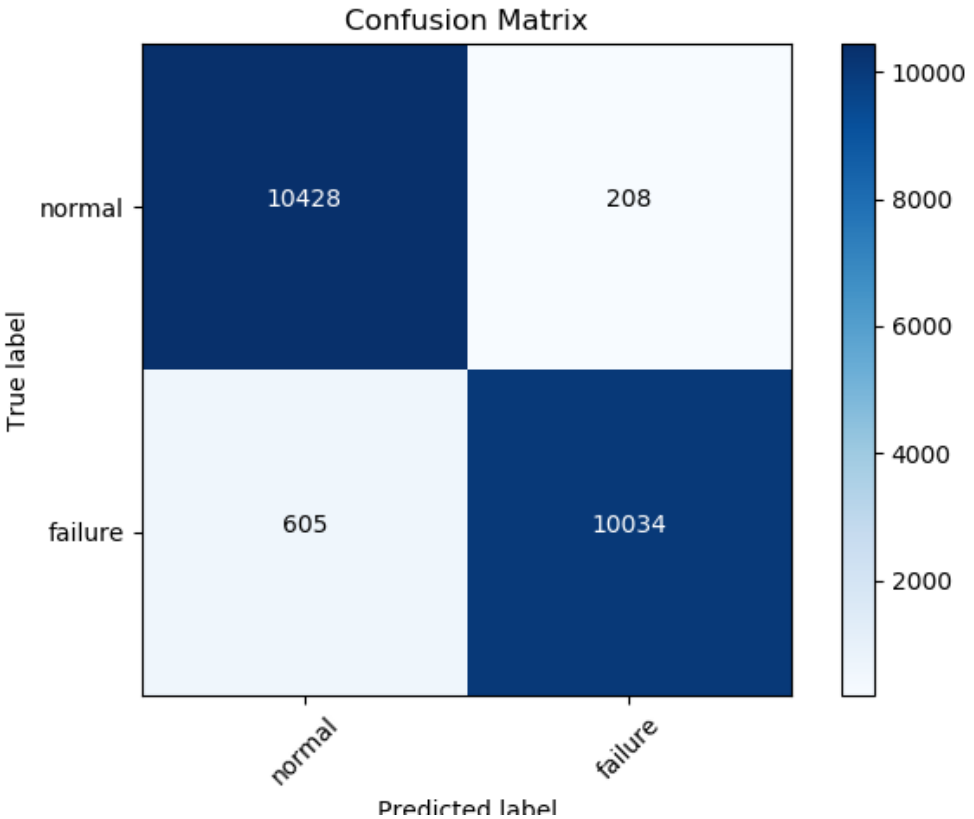


图 4.3.2(b)　基于过采样数据集的 Random Forest 算法结果

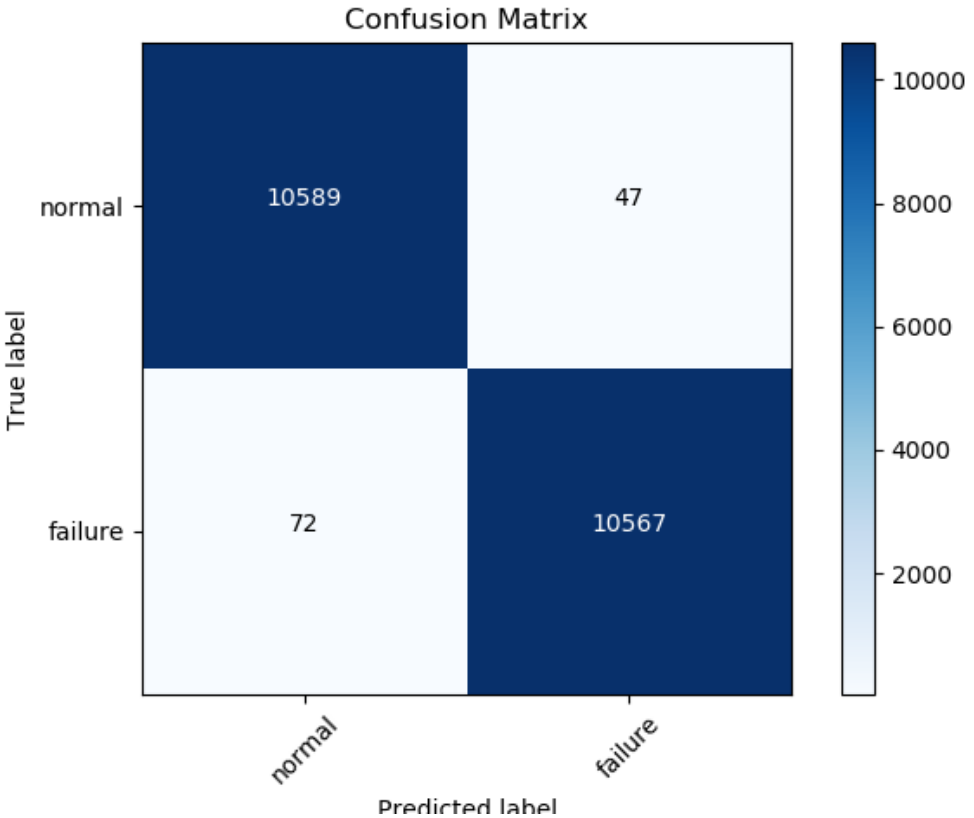


图 4.3.2(c)　基于过采样+降采样数据集的 Random Forest 算法结果

4.4 时序预测分类算法——卷积神经网络

4.4.1 算法基本原理

卷积神经网络本质上是一种从输入到输出的映射网络，该方法能够规避精确的数学解析表达式，其通过学习大量输入输出间的映射关系，利用已知的模式对卷积神经网络加以训练，神经网络便能够具有输入输出间的映射能力。卷积神经网络执行的是有监督训练，其样本集是由形如：（输入向量，理想输出向量）的向量对构成。在训练开始前，权重系数与偏置均采用较小的随机数生成。小随机数用来保证神经网络不会因权值过大而进入饱和状态，从而导致训练失败，不同的权值以保证神经网络能够正常地学习。实际上，若是使用相同的权值去初始化矩阵，则神经网络将丧失学习能力。

卷积神经网络包含卷积层、池化层与全连接层。卷积层的意义在于特征提取与权值共享，某一特定数据点与局部周围的数据点关系紧密，而和较远距离的全局像素关系不是非常密切，卷积层通过卷积核 filter 将邻近区域内的数据点卷积在一起，从而能够提取出较好的局部特征，通过多层卷积，能够将卷积的范围进一步扩大，从而使得特征具有全局意义。而权值共享则降低了网络的复杂度，避免了特征提取和分类过程中模型重建开销。权值共享的实现是通过将图像的一块（局部感受野）作为层级结构中最低层的输入，信号再依次传输到不同的层，每层通过一个数字滤波器（或探测器）去获取感知数据的最显著的特征。

池化层的加入能够有效地降低运算数据量，其意义在于降维、非线性实现、扩大感受野以及实现不变性，进而减少神经网络的数据负载，加快模型的训练速度。卷积神经网络特征检测层通过数据训练进行学习，隐式地从训练数据中进行深度学习，从而避免显式的特征提取。同一特征映射面上的神经元权值相同，意味着可以进行并行训练学习。流行的分类方式多是基于统计特征，这意味着在进行分类前必须提取某些特征。

图 4.4.1　卷积神经网络预测模型示意图

4.4.2 基于时间滑窗的数据选取

由于风机结冰过程不是一蹴而就的，风机结冰问题具有较强的连续时间趋势关系，因此，需要综合考虑 28 维变量和时间信息对分类结果的影响。

基于上述考虑，对于卷积神经网络的输入数据，不妨进行滑窗选取的操作。将某一条数据及其之前的 63 条数据作为 CNN 的输入而 CNN 的输出选为该数据所对应的“正常/结冰”标签，进行卷积神经网络模型的训练。

关于“时间滑窗”的具体操作，采用固定的步长滑动选取一定窗宽数据的方式，在这里选取了步长为 1、窗宽为 64 的方案。

这是因为窗宽的大小会影响模型预测的准确度：过大的窗宽虽然包含了更多的时间维度信息，但会不可避免地造成数据处理量的增加，不利于算法效率的提高；而过小的窗宽会造成时间维度信息不足的缺陷，不能够很好地体现连续时间信息对于风机结冰情况预测的重要作用。

4.4.3 最终结果与对比

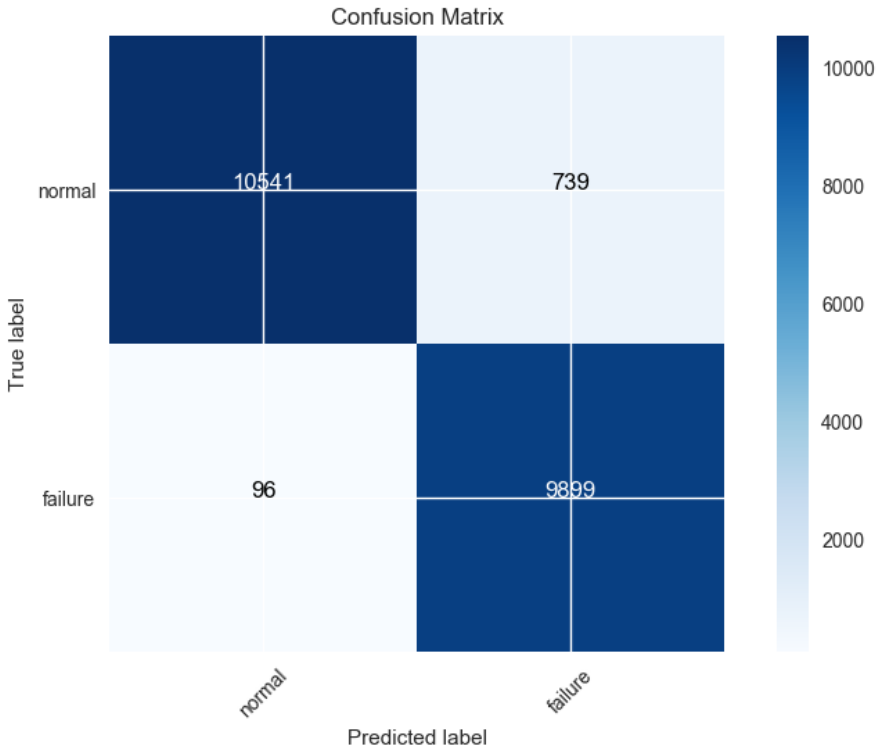


图 4.4.3(a)　基于原始数据集的卷积神经网络算法结果

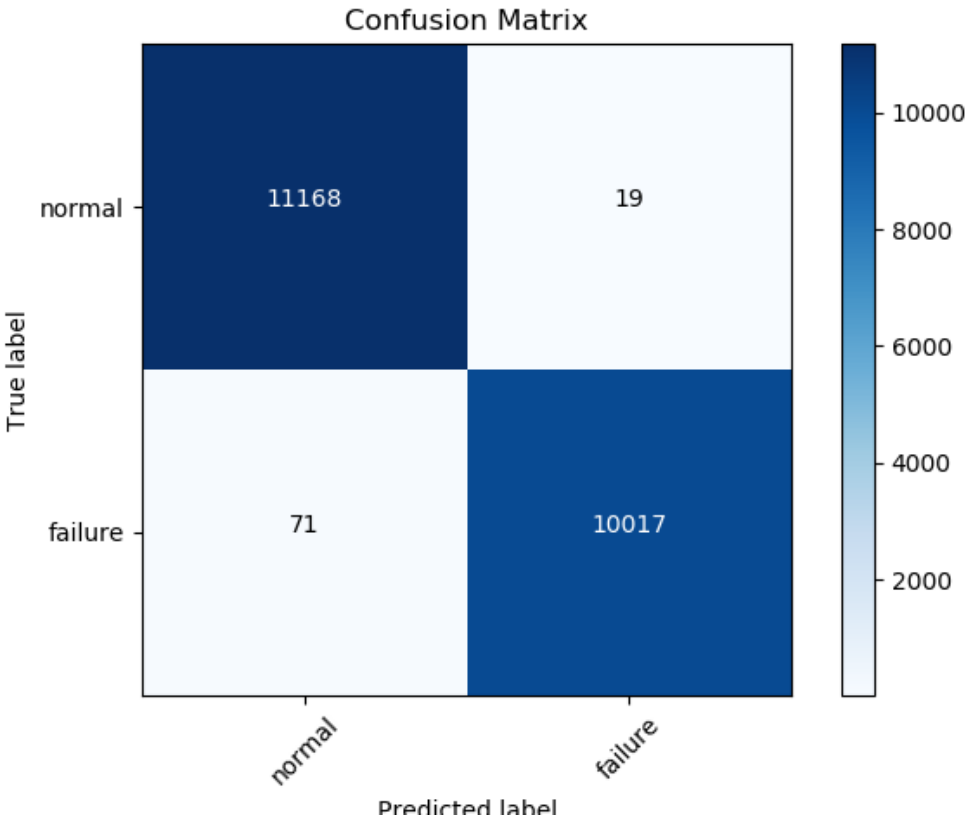


图 4.4.3(b)　基于过采样数据集的卷积神经网络算法结果

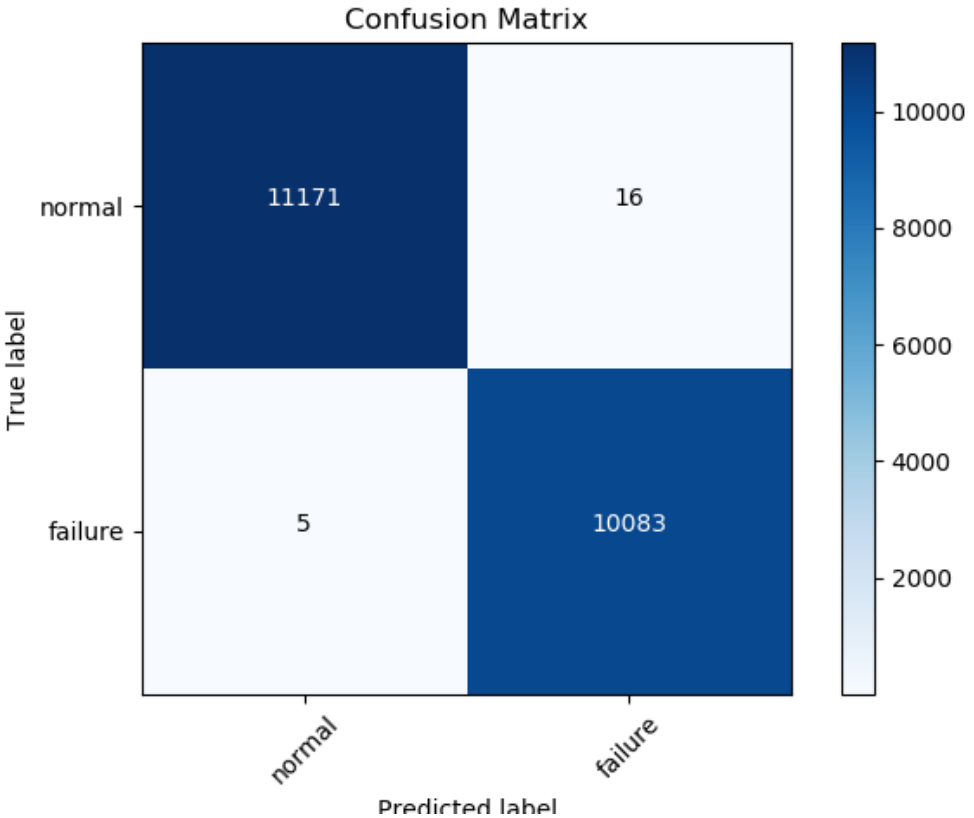


图 4.4.3(c)　基于过采样+降采样数据集的卷积神经网络算法结果

4.5 预测融合算法——基于 CNN 的白天黑夜模型

4.5.1 算法基本原理

基于卷积神经网络的时序分类算法，其预测正确率很大程度上依赖于训练数据量的大小。希望减小训练数据量的大小，一方面可以通过对原始数据集进行数据预处理来完成，另一方面也可以通过优化模型结构来达到较少训练数据量的目标。本问题求解中，考虑到白天与黑夜不同时间段风机结冰情况的差异，我们采用预测融合的方法，以达到能够更具针对性地处理数据的目的。

问题求解过程中，我们共构建了三个子卷积神经网络。第一个卷积神经网络依据数据特征，对所处时间段是白天或黑夜进行判断；第二与第三两个卷积神经网络分别为基于白天数据与基于黑夜数据训练的卷积神经网络模型。数据首先通过第一个卷积神经网络进行白天黑夜判断，依据判断结果，再将数据分别对应送入白天或黑夜的子预测模型中，以更具针对性地对数据结果进行预测。

白天黑夜预测模型的结构示意图如下图所示：

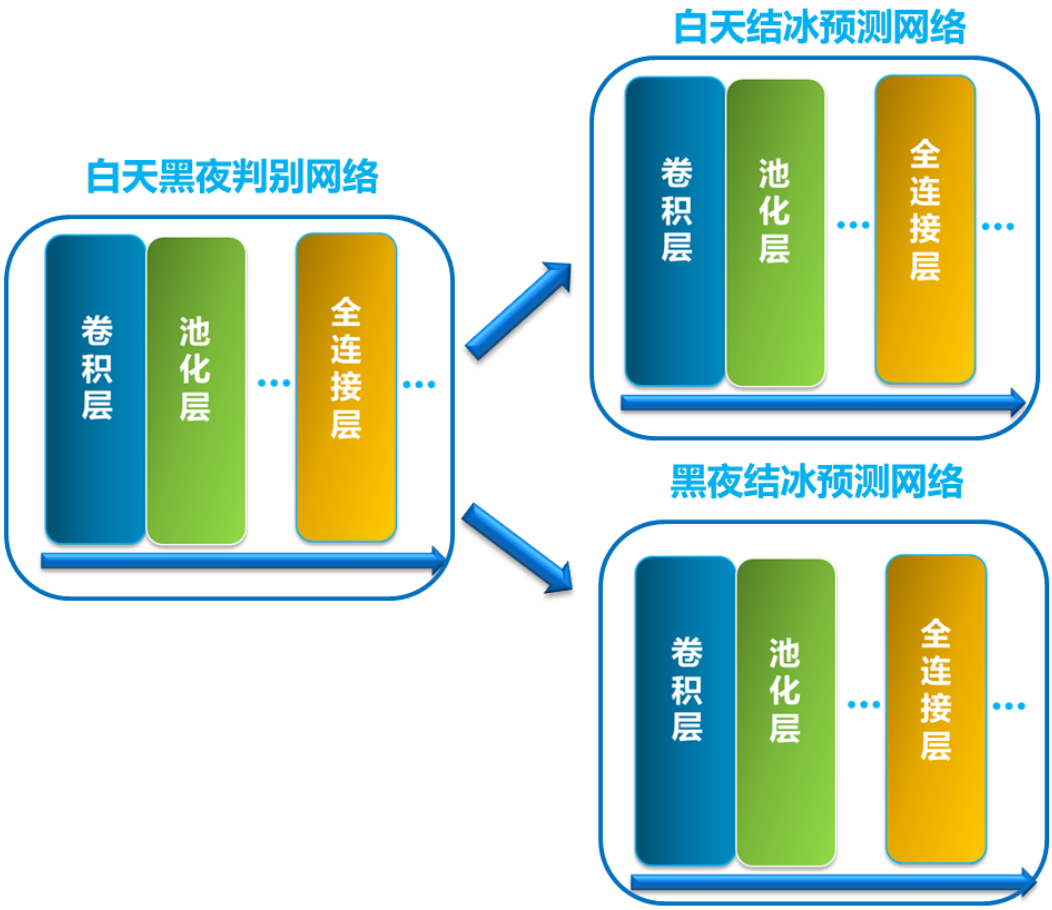


图 4.5.1 “白天-黑夜”卷积神经网络模型结构示意图

4.5.2 最终结果与对比

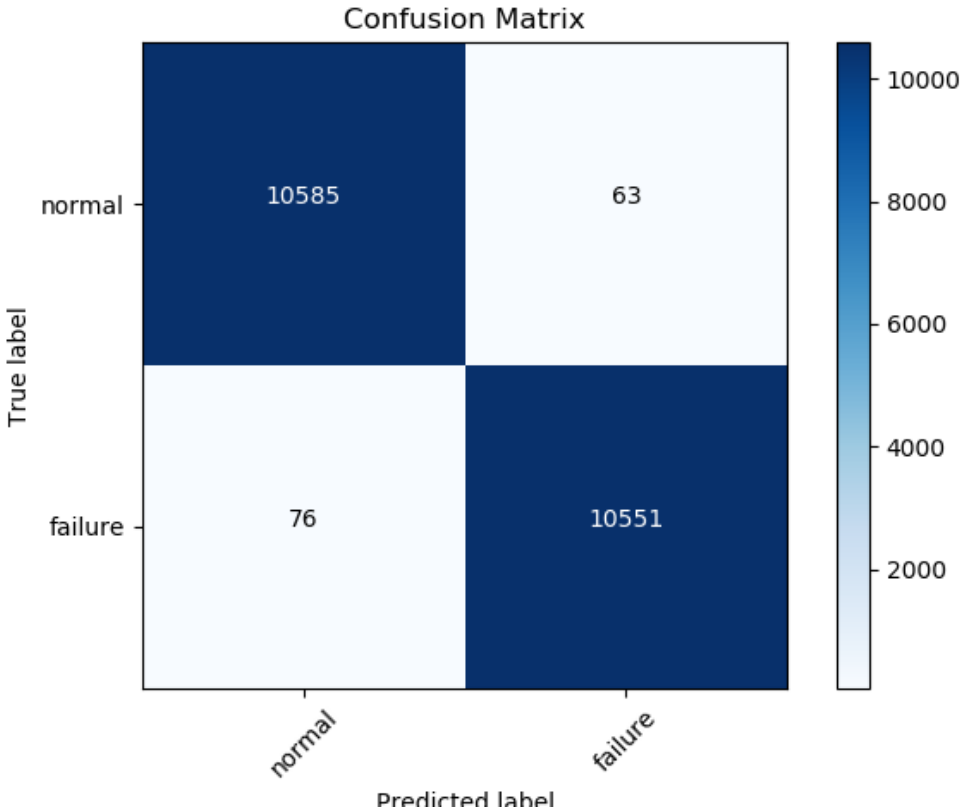


图 4.5.2(a)　基于原始数据集的白天黑夜模型结果

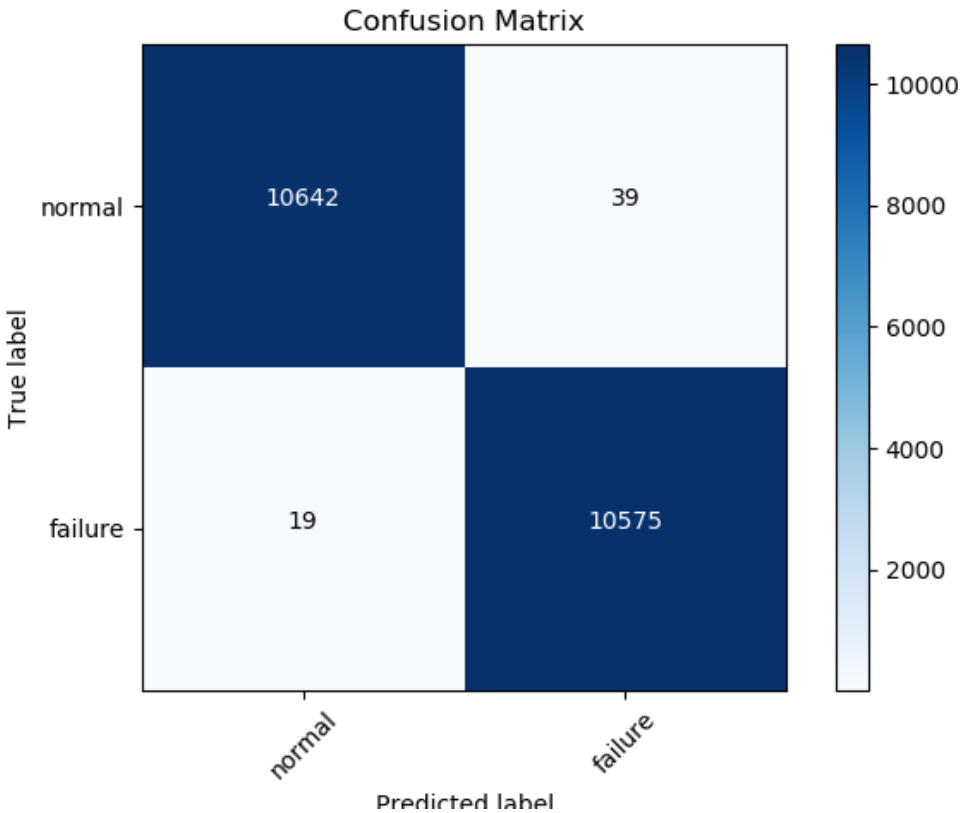


图 4.5.2(b)　基于过采样数据集的白天黑夜模型结果

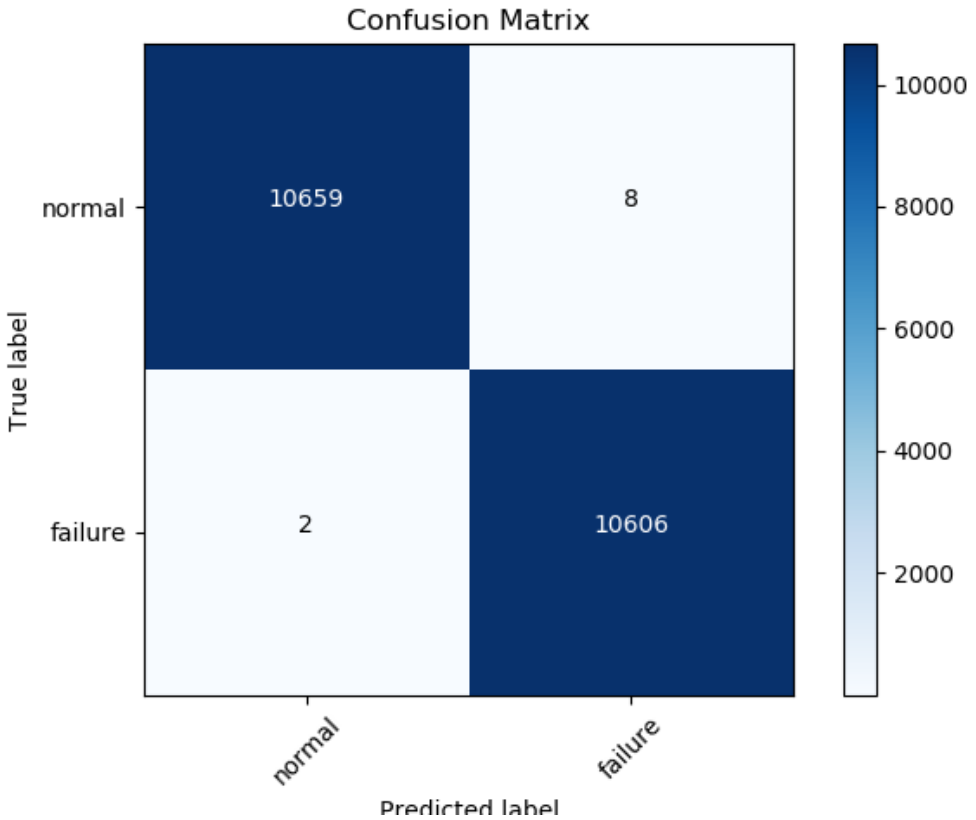


图 4.5.2(c) 基于过采样+降采样数据集的白天黑夜模型结果

5 模型数据对比与评估

基于不同运行数据集运行上述四种模型，记录不同数据集、不同模型下的模型准确率及运行时间，结果展示如下表所示。

	原始 (40w)	过采样 (80w)	过采样+降采样 (30w)
Fisher 判别	52.13%	51.18%	51.83%
随机森林	93.83%	96.18%	99.44%
CNN	96.08%	99.58%	99.90%
CNN 白天黑夜模型	99.35%	99.73%	99.95%

通过表格可以直观看到基于不同数据、采用不同算法的预测结果。纵向比较表格，以评估不同算法在预测问题中的表现。我们可以明显看出，采用以 Fisher 判别为代表线性分类器其预测准确度较差；而采用以随机森林、CNN 为代表的非线性分类器在该预测问题中表现优异。随着过程中我们对模型的进一步改善，可以看到，在非线性分类器中，以基于 CNN 的白天黑夜模型表现最为优异，这样的结果也是与我们的预期相符的。

横向比较该表格，以评估不同数据处理结果对预测问题准确性的影响。通过平衡数据集，我们能够有效地预防预测模型过拟合的情况，从而提升模型的泛化能力与准确性。我们可以明显看出，通过对模型进行过采样、降采样操作，以改善类间数据的不平衡性，对于模型的预测结果有着较好的改善效果，这样的结果也是与我们的预期相符的。

6 模型优化与后续工作

6.1 数据清洗的改进与提升措施

✧ 采用 K-means 聚类方法筛查离群点：现阶段我们采用目测的方式观察数据分布，手动设置阈值，删除数据，具有较强的主观性，使用聚类的方式能提供一个更加统一的标准，使得离群点数据的筛查更加具有科学性。

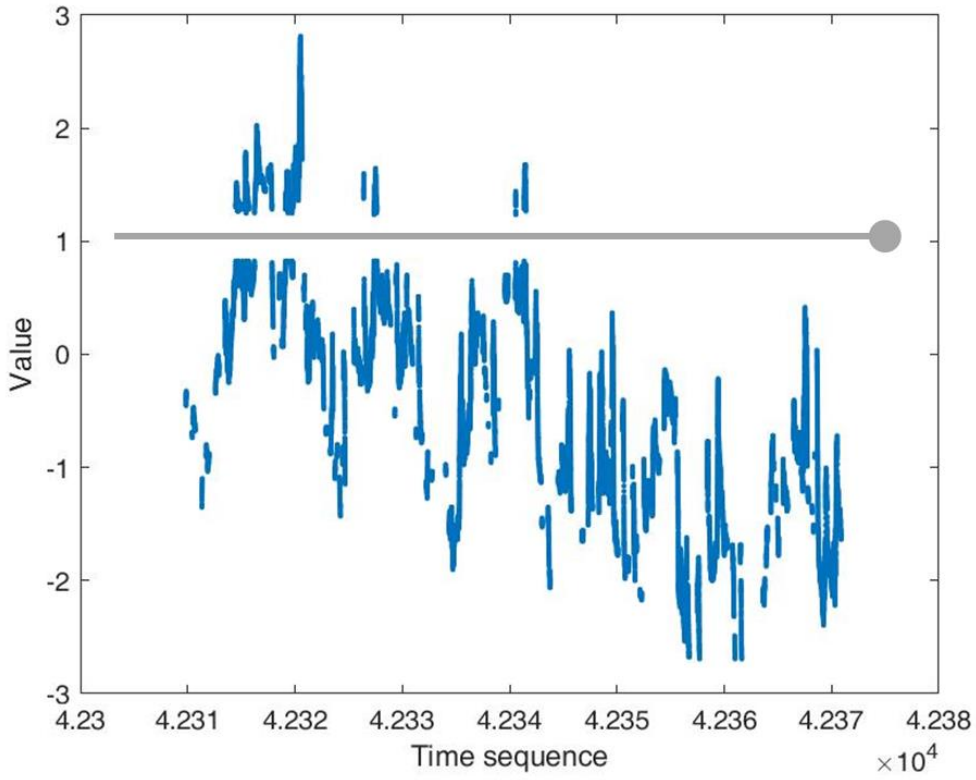


图 6.1　原始数据集“环境温度-时间”示意图

✧ 对人工删除的数据进行插值处理：在绘制“变量-时间”图的过程中，我们可以明显地观察到部分数据存在整体删除的情况，例如在“温度-时间”图中，1℃附近的温度被人为删除了。这种删除带来的，是整行数据的缺省，对数据的连续性带来了很大的干扰，因此插值为连续数据是很必要的。然而，关于插值的方式上，因为缺少一种具有说服力的插值方式，目

前仍有待讨论，可以考虑针对不同插值方式进行比较分析。

- ✧ 连续的数据离散化：通过先验知识，将数据分集。例如，可以将温度分成零上和零下部分。这样做的目的实则是在人为地设置权重，加强同一特征不同大小间的差异。在有先验知识保证的情况下，可以提高模型的准确性，提高泛化能力。
- ✧ 利用强规则进行过滤：最终的目的是要预测错误数据，因此对于一些明显正确的数据，我们可以选择性地删除，从而降低样本数量，提高模型鲁棒性。

6.2 特征工程的改进与提升措施

- ✧ 构造新的物理变量：进一步地构造其他表征结冰状况的物理变量，提高模型准确性。

6.3 数据建模的改进与提升措施

- ✧ 进一步优化卷积神经网络结构。
- ✧ 构建多分类器融合的分类器：综合考虑模型的线性关系，非线性关系，时序关系，通过权重的设置，获得泛化能力更强的模型。

参考文献

[1] Z. Zhang, Zhe Song and J. Xu, 2015, Data-Driven Wind Turbine Power Generation Performance Monitoring, IEEE

[2] Transactions on Industrial Electronics,Vol.62。Lee, S., Park, W., and Jung, S. (2014). Fault detection of aircraft system with random forest algorithm and similarity measure. The Scientific World Journal, 2014.

[3] 彭深. 基于综合指标的叶片结冰监测方法[A]. 中国农业机械工业协会风力机械分会.第四届中国风电后市场专题研讨会论文集[C].中国农业机械工业协会风力机械分会:,2017:4

[4] 东乔天, 金哲岩, 杨志刚. 风力机结冰问题研究综述[J]. 机械设计与制造, 2014(10):269-272.

[5] 工业大数据创新竞赛白皮书(2017)