工业大数据创新竞赛白皮书

（2017）

风机结冰故障分析指南

指导单位：工信部信息化和软件服务业司

工业互联网产业联盟

编写单位：工业大数据创新竞赛组委会

2018 年 1 月

编写说明

2017 年的工业大数据竞赛作为我国首次工业大数据竞赛，在吸引人才关注、促进工业智能化、建立工业大数据生态等方面起到重要作用。习近平总书记在党的十九大报告中强调:“建设现代化经济体系，深化供给侧结构性改革，加快发展先进制造业，推动互联网、大数据、人工智能和实体经济深度融合。”这些都为制造业转型升级指明了新方向，数据成为制造业与新一代信息技术融合的重要基础资源和创新引擎。在工信部信息化和软件服务业司、工业互联网产业联盟指导下，本次竞赛组委会在北京天泽智云科技有限公司的倾力支持下，组织参赛者编写了《工业大数据竞赛白皮书（风机结冰故障分析指南）》，希望将本次竞赛的经验与技术成果固化并加以推广，促进交流，与业界共同推动工业大数据发展。

白皮书收录了 2017 年工业大数据竞赛-风机叶片结冰故障预测的获奖算法，组成解法集，在工业大数据分析的方法论上具有重要的指导意义，在风力发电机行业尤其具有示范作用。白皮书主要分为四个部分，第一部分是工业大数据创新竞赛概况，主要介绍此次竞赛背景和开展情况。第二部分提出了工业大数据分析方法论，并以本次数据竞赛的题目作为案例，解释预测性建模的各个分析流程。第三部分主要收录竞赛优秀算法，包括文献调研、方法介绍、方法应用与验证、与结果讨论。第

四部分进行方法论总结。

白皮书的编写过程中得到竞赛组委会和参赛选手的大力支持。相关参赛人员根据自己对竞赛题目的解法，给出了详细、清晰的方法流程与结果讨论，为白皮书的提供了丰富的素材与扎实的内容。同时，不同解法也为工业大数据分析从业人员打开思路、拓宽视野提供了极具价值的参考。

工业大数据的发展还在初级阶段，随着工业大数据竞赛日后的逐年举办与工业智能的逐步发展，我们将根据各界的反馈意见，在持续调研与总结的基础上，定期进行修订与新版发布。

指导单位：工业和信息化部信息化和软件服务业司

工业互联网产业联盟

编写单位：工业大数据创新竞赛组委会

指导专家：

孙家广 中国工程院院士

谢少锋 工业和信息化部信息化和软件服务业司司长

安筱鹏 工业和信息化部信息化和软件服务业司副司长

李 杰 美国辛辛那提大学智能维护系统（IMS）中心主任

王建民 清华大学软件学院院长

林诗万 IIC 技术工作组与架构任务组联执主席

余晓晖 中国信息通信研究院总工

编写组成员（排名不分先后）：

工业和信息化部信息化和软件服务业司：王建伟、冯伟

中国信息通信研究院：冯旭、朱敏、刘默、李铮、宋辰超、李南、魏凯

北京天泽智云科技有限公司：金超、晋文静、刘宗长、李飞

北京工业大数据创新中心：王晨、田春华、崔鹏飞

金风科技：张光磊、周杰、李富荣

万腾科技：张嗣昌、侯振寰、侯宗波

济中节能：纪浩然、张薇、王成金

浙江运达风电股份有限公司：周书锋、朱博文、冯文婷

北京邮电大学：林文芳

西安交通大学：李宁波、闫涛、郭亮

序言

随着新工业革命时代的序幕徐徐拉开，物联网、工业互联网、智能 ICT 技术、人工智能等技术成为舞台上最受瞩目的新星。在这些新兴技术的推动下，工业领域中的大数据环境正在逐渐形成，数据从制造过程中的副产品转变成为备受企业关注的战略资源，成为工业企业传承制造知识和提供增值服务的依托。然而，工业大数据在其可获取性和可分析性方面仍然存在许多的挑战，一方面企业拥有大量数据但缺乏专业的数据分析人才，而另一方面拥有分析能力的人才缺少数据和应用场景。工业大数据由于其应用场景的专业性与多样性，使之兼具工业体系的系统性与互联网的开放性，也使企业很难独立建立完整的工业大数据应用能力。在这样的挑战下，需要建立一个开放的生态，将数据和场景的提供者、知识和能力的提供者、产业链相关上下游聚合在一起，让数据的生态、知识的生态和服务的生态得以相互促进。

建立可持续的大数据人才培养模式和人才培养体系是产、学界面临的共同挑战。可持续的人才培养模式不仅局限于大学中，还包括企业内的人才培养。大学的人才的培养需要鼓励创新性和独特性，而企业则应该注重员工专业技能和应用研发能力的培养。工业大数据竞赛是非常好的产、学界共同携手培养人才的方式。产业界贡献场景和数据，可以帮助学术界的研究

更贴近真实需求。而学术界也为企业提供最新的理论和最前沿的技术成果，拓宽了企业解决问题的视野。美国在这方面的投入已经持续了多年，从 2008 年开始美国的 PHM 学会（PHM Society）就开始举办工业数据分析竞赛。数据的贡献者主要来自于企业或产业研究机构，涉及的行业非常广泛，但都遵循着同一个原则，就是场景都来自于企业的真实问题，数据都来自于真实的工业现场。这个竞赛中所使用的数据可供全世界的研究者下载，比赛的胜出者也会受邀在其期刊中发表论文共享好的分析方法。IMS 中心参加了从 2008 年至今的 10 次数据竞赛，获得了其中的 5 次冠军，所贡献的方法在工业界中得到广泛应用。

在本届工业大数据竞赛中，我们欣喜地看到参赛队伍包括了产业界和学术界，参赛的企业包括风电装备制造、风场运营商、服务提供商、以及其他工业领域的企业，总数超过了 1000 多只参赛队伍。本次数据竞赛在引领和催化工业大数据应用生态形成方面的作用是有目共睹的。

今年 IMS 中心有幸作为顾问单位参与中国第一届工业大数据竞赛的组织工作，竞赛获得的关注程度以及选手们在竞赛中的表现都令人感到惊喜。《工业大数据创新竞赛（2017）白皮书》作为本次竞赛的重要成果之一，对工业大数据分析方法论进行了系统性地介绍，并对竞赛优胜团队的解题方法进行了详细地整理和解读，相信能够为从事工业大数据应用研究的企业和学者们提供有价值的参考。衷心祝愿工业大数据创新竞赛越办越

好，成为产学界共同推崇的传统和品牌，为中国工业大数据产业生态源源不断地输送优秀人才。

李杰，2018 年 1 月

目 录

一、工业大数据创新竞赛概况

（一）数据经济的崛起与工业的变革

当前，**世界经济加速向以网络信息技术产业为重要内容的经济活动转变，数字经济正深刻地改变着人类的生产和生活方式，成为经济增长新动能。**人类社会正在被网络化连接、数据化描绘、融合化发展，在这一进程中，数据成为重要的基础性战略资源。大数据的充分挖掘和利用，极大促进了全社会要素资源的网络化共享、集约化整合、协作化开发、高效化利用，对经济发展、社会生活和国家治理产生着越来越重要的作用，推动了诸多领域发生重大而深刻的变革，一个全新的大数据时代正在向我们大踏步地走来。

大数据是一种资源，一种技术，一种产业，更是一个时代，它通过构筑信息互通、资源共享、能力协同、开放合作的发展新体系，为提升政府治理能力、优化民生公共服务、促进经济转型和创新发展做出了积极贡献。尤其是随着近年来，互联网产业对数据价值挖掘的成功，使得传统行业开始思考如何推动价值转型，驱动工业变革，这是新的技术条件下制造业生产全流程、全产业链、产品全生命周期的数据可获取、可分析、可执行的必然结果，也是制造业隐性知识显性化不断取得突破的内在要求。

在数据经济浪潮的推动下，是否能对数据进行深度的价值挖掘，将是各个行业竞争的新重点。值得明确提出的是，数据

本身并不能创造价值。如果只是对数据进行收集、存储、与管理，是无法为工业飞跃式地创造价值的 。为了实现工业大数据驱动的价值转型，需要从工业中的问题出发，将业务问题转化为数据预测性建模问题，从而达到解决用户痛点、实现用户价值转型的目的。

（二）工业大数据驱动制造业转型升级

大数据作为一种新的资产、资源和生产要素，正驱动着制造业的智能化变革，可以从三方面来理解。首先，资源优化是目标，工业大数据的创新价值集中体现在制造资源配置效率的优化，以及制造业全要素生产率的提高。**其次，信息流动是关键，**工业大数据如何优化制造资源配置效率，关键是要把正确的信息在正确的时间传递给正确的人和机器，解决制造过程的复杂性和不确定性等问题。**第三，大数据、人工智能、互联网等新一代信息技术是基础，**为数据的全面感知、在线汇聚和智能分析构筑赋能工具和载体，这正是工业大数据的核心功能。

关于大数据的分析方法，人们首先想到的可能是 Hadoop，Spark 等 IT 技术。然而，对于工业中的大数据问题，其重要价值在于形成并不断优化认识和改造世界的方法论，除了分析平台与数据处理基础设施，用户应该更加关心大数据分析所能带来的价值，再选择与分析目标相适应的技术。

通过工业大数据创造价值，需要围绕业务目标，将基于机理模型的模拟择优法和数据模型驱动的大数据分析法进行融合。正如在首届（2017）工业大数据创新竞赛决赛答辩和颁奖仪式

上，工业和信息化部信息化和软件服务业副司长安筱鹏所指出的：机理模型与数据模型的融合，能够突破隐性数据显性化和隐性知识显性化两大关键，通过构建制造业快速迭代、持续优化、数据驱动的新方式，解决发生了什么、为什么发生、下一步发生什么、如何改进优化四个问题，优化制造资源的配置效率。

通过工业大数据创造价值，需要整合不同学科、不同领域的经验、知识和技术。美国国家自然基金会产学合作智能维护系统中心主任李杰教授提到，工业大数据遇到的挑战主要可以分为工业场景的复杂性与不确定性两个方面。在复杂性方面，从数据接入、数据治理、模型建立、优化分析、到最后的决策支持与行动，价值转型的过程就是打通工业大数据信息与知识链路的过程，关键在于如何融合数据技术、分析技术、与运营管理技术。在不确定性方面，工业大数据分析的不确定性体现在应用对象、工况、数据、模型等等诸多因素上。

通过工业大数据创造价值，也需要面对工业应用对象差异所带来的挑战。清华大学软件学院院长王建民强调了数据模型泛化的能力，在物联网数据采集规模化的今天，如何能够把训练好的模型快速准确的应用到另一个同类对象上，是工业大数据实施过程中的重点。同时，中国工程院院士孙家广也阐述了工业大数据价值创造处在起步的初级阶段，而本次创新竞赛是推进工业大数据普世化、规范化、国际化、市场化的绝佳机会，要坚持走出有中国特色的工业大数据技术与产业创新道路，助

理中国工业由大变强、弯道超车。

工业大数据并不单单是某一种技术，而是一种理念。企业若想要通过工业大数据实现价值转型，需要打破原有的技术采购思路，不断提升技术和组织能力，从自身问题和需求出发，用工业大数据的方法论切实解决问题和创造价值。而一个行业要想通过工业大数据实现产业升级，则需要更加开放的生态、与共赢的思维，对行业通用的痛点进行充分地讨论与系统性的梳理，为实现行业的平台化、标准化、规模化工业大数据环境奠定基础。

（三）工业大数据创新竞赛开展情况

为进一步探索工业大数据对工业改革的深远影响，由工业和信息化部指导，在工业和信息化部指导下，以“赋能与赋智，构建工业大数据应用生态”为主题，以“开放共享、协作共赢”为原则，中国信息通信研究院联合业界同仁举办首届工业大数据创新竞赛，这也是首次由政府主管部门组织的工业大数据领域权威的全国性创新竞赛。大赛在发掘专业技术人才的同时，助力于解决工业企业实际问题，以提升制造智能水平，推动中国工业转型升级，推进工业大数据的加速发展，积极促进赛事成果转化和产学研用紧密结合，服务工业经济提质增效升级，推荐优秀专业技术人才找到适合发展的平台。

本次比赛围绕风电装备预测性维护这一应用场景，针对风机叶片结冰故障预测和风机齿形带故障两个真实工业大数据应用需求，由金风科技分别提供来自于某风场的 13 台风机半年运

行数据，每台风机包括工况、环境等 28 个变量，设置两道赛题面向全社会征集解决方案，旨在通过竞赛方式解决大数据技术在工业应用落地过程中面临的“有数据没技术、有技术没应用场景”等问题。竞赛过程中，得到了北京工业大数据创新中心、树根互联技术有限公司、星河互联、美国国家仪器、美国国家科学基金会智能维护系统中心等企业和研究机构的大力支持。

活动自 2017 年 7 月启动，至 2017 年 12 月正式结束。**竞赛注册用户数 1535 人，分别有 830 支和 630 支队伍参加两个竞赛题目，其中 60%以上来自于高校学生，涉及数据挖掘、控制工程、工业机器人、测控技术、计算机等多个领域。**竞赛分初赛、复赛和决赛三个阶段进行，经过 3 个月的角逐和复赛专家评审，最终 12 支队伍获奖，其中 3 支队伍是北邮、西安交大和浙大的学生团队，其余 9 支是企业团队，包括富士康等制造企业，以及信息通信、能源等领域的初创企业。

附：获奖队伍

比赛一：风机叶片结冰预测		比赛二：风机齿形带故障分类	
一等奖	万腾科技（团队）	一等奖	中国石油中油瑞飞（AC_Drilling 团队）
二等奖	济中节能（团队） 运达风电（世属三团队）	二等奖	浙江大学（DCL 团队） 南京大学（Diaryfly 团队）
三等奖	富士康科技（个人） 西安交通大学（XJTU_DL 团队） 北京邮电大学（个人）	三等奖	难愚科技（团队） 富士康科技（Knight 团队） 富士康科技（DPBG_IT 团队）

二、工业智能分析方法论

（一）工业智能分析方法流程

简单来说，工业智能指的是人工智能技术在工业中的应用。工业智能的萌芽得益于人工智能技术的进步，其技术驱动因素包括传感器成本的降低、计算能力的飞跃和机器学习算法准确度质的提升。在多个通用人工智能领域，人工智能算法的准确性都取得了巨大突破。而传统的工业生产活动中，直至今天依然非常依赖人力、经验、与设备本身，用户往往重视管控“可见问题”，而忽略了挖掘“不可见问题”。在工业对象数据的自动化获取与存储越来越容易越来越廉价的今天，工业作为智能化程度的“洼地”及其潜在的巨大商业价值，受到的关注与日俱增。

工业智能并不是通用人工智能技术在工业场景中的简单复用。工业场景中问题的碎片化、个性化、与专业化的特点，决定了工业智能落地需要依靠计算机科学、人工智能、与领域知识的深度融合。**与传统基于规则或单纯依赖机理建模的方式不同，数据驱动的工业智能技术的一大优势是通过基于统计意义上的预测性分析，对不确定性更加有效地管理，同时更好地结合专家知识并将其固化到软件中，形成可持续迭代的智能系统。**

工业系统的智能化转型主要主要体现在以下三方面[1]：

一是从基于经验的决策转变为基于实证的决策：传统的工

[1] Lee, J., Bagheri, B., & Jin, C. (2016). Introduction to cyber manufacturing. *Manufacturing Letters*, *8*, 11-15.

业系统高度依赖专家的经验。随着专家年龄的增大、员工离职率的逐渐攀升，这种经验越来越难以传承。企业为了可持续发展，专家的经验需要以某种方式固化下来成为模型、判断标准、流程等，支持企业中的各个方面在正确的时间做出正确的决策。

二是从解决可见问题转变为避免不可见问题：工业中的问题可以被分为可见与不可见两类。工业活动中的经典管理与分析策略，绝大多数都聚焦在解决可见问题，如设备定期维护保养，产品质量抽检，机器换人等。这些方法的应用并不能阻止出现设备的非预期停机、不良品的出现与根因分析的困难、以及自动化机器误操作等问题。这是由于生产中如设备关键组件衰退、工艺过程与质量关系等不可见的问题没有被量化。将隐性问题与隐性关系显性化，才能够从根本上帮助用户降低成本、提升效率。

三是从基于控制的自动化转变为基于机器学习的智能化：自动化系统曾被认为是工业智能的核心。然而，单纯的自动化仍无法完全满足工业智能化的需求。自动化系统能够解决的是能够被相对清晰、明确定义的问题，即可见问题；而智能化系统要暴露的往往是可见问题暴露之前的隐性问题。同时，在现代工业系统变得越来越复杂的情况下，人工智能算法能够与自动化模型相结合，超越传统控制的局限性，实现全局优化，达到增强系统强健性的目的。

在工业系统中，设备的预测性智能维护和效能动态优化是工业大数据的核心应用场景之一，也是实现智能化工业系统最

为关键的核心技术之一。对设备性能的预测分析和对故障时间的精准估计，将量化管理设备运行中的不确定性，并减少这些不确定性的影响，来为用户提供预先缓和措施和解决对策，以防止设备运行中的非预期停机损失和事故风险。同时，根据设备的健康状态、外部环境、产线组织形式和生产目标等多维信息，基于工业大数据的预测性模型可以对产线整体的效能进行优化决策支持，从而实现对生产系统成本和效益的深度管理和效益提升。

数据驱动（Data-driven）的分析手段并非是对设备的状态和效能进行建模和预测的唯一途径，其他的方式还包括物理建模、可靠性模型、和混合模型等。在对数据驱动方法的原理进行阐述之前，首先要对‘特征’这个重要概念进行解释。

特征的含义是，从数据当中抽象提取出的与判断某一事物的状态或属性有较强关联的可被量化的指标。例如在人脸识别的过程中首先要提取出脸部主要器官的位置、形状、相对距离等特征，再对这些特征进行匹配，从而实现身份的识别。在设备健康状态预测方面，提取有效的健康特征对预测准确性至关重要。常用的特征包括时域信号的统计特征、波形信号的频域特征、能量谱特征、特定工况下的信号读数等。

在对原始数据进行特征提取之后，智能算法通过对多维度数据的融合分析来建立健康预测模型。基于统计或机器学习的算法能够根据所定义的目标函数来优化预测性模型的结构与参数，从而“记忆”数据中的信息，并能据此对类似的数据做判

断。模型所记录的信息可以是多维数据的模态，其与某一状态的相似度，或者是输入特征之间的相关性等。以对设备的状态评估为例，图 2-1 的横轴与纵轴分别代表两个不同的特征，在两个特征构成的特征空间中，设备的不同健康状态对应着特征的不同分布。在制造系统的运行过程中，随着设备衰退，其对应的特征分布会慢慢偏移。特征的分布与正常状态分布的重叠部分表征设备当前状态与健康状态的相似度，即其健康的可能性，或称之为“健康值”。若设备的故障状态特征分布已知，则该健康值将可以被正态化为 0-1 之间的量。随着时间的推移，这个分布会慢慢向某一个失效状态发展，j 对应的健康值时间序列代表的是设备衰退的轨迹。如果进一步对这个趋势的发展进行预测，就可以推断出在未来的什么时间会发生什么问题或故障。

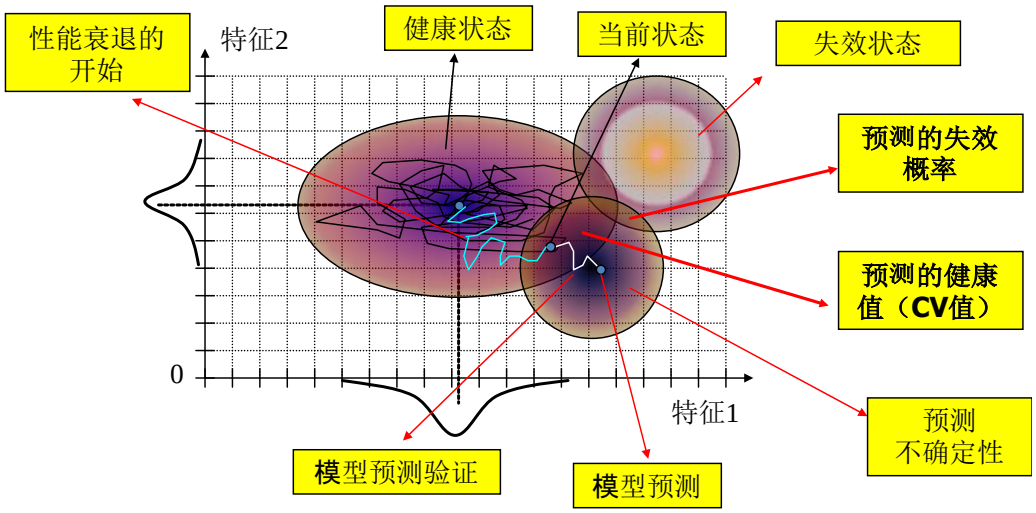


图 2-1：利用大数据建模分析制造系统隐性问题的原理

[来源：美国智能维护系统中心]

对工业大数据的建模，其目的是为用户创造价值。这决定

了工业大数据的建模过程需要以业务目标为驱动，同时要融合算法科学、领域知识、与软件工程。如图 2-2 所示，工业大数据的分析过程包括了业务场景分析、数据问题定义、数据场景化、模型建立、模型价值评估、以及最后的部署实施六大步骤。

第一步，业务场景分析：工业大数据分析不同于互联网，对通过数据挖掘来泛泛寻找相关性这种模式在成本上无法承受。工业大数据分析应该从业务入手，在了解行业背景、分析用户痛点之后，制定明确的数据服务目标，定义工业数据分析系统的功能与边界。

第二步，数据问题定义：在确定业务目标之后，需要对问题进行数学化的定义。在工业中，并非所有的问题都适用于数据驱动的建模方式。根据数据的数量、质量、与可采集变量的完整性，明确数据建模的策略与详细流程。

第三步，数据场景化：原始数据往往因为数据质量、工况完整性、标签缺失等问题无法用来直接建模。在建模之前，有必要检测数据质量，将数据与业务场景相对应，之后提取能够反映建模对象健康状态的特征，为后续模型输入做准备。

第四步，模型建立：这一步与通常意义上的机器学习过程类似。不同的是，在工业数据预测性分析中，建模是更加强调模型的可靠性、泛化能力、以及可解释性。

第五步，模型价值评估：模型本身性能与准确性不是工业数据分析的唯一衡量标准。如何能够让模型产生准确的可执行信息，快速支持用户决策，改善设备健康状态，优化运维效率，

是建模中需要着重强调的关键评估角度。

第六步，模型部署与实施：模型本身不产生价值，嵌入软件产品中支持业务改善的模型才有价值。与离线的验证不同，工业系统的模型上线后，仍需要被维护、管理、以及不断迭代，以适应变换的工业场景与可能出现的问题，持续为用户提供设备洞察，提高生产力。

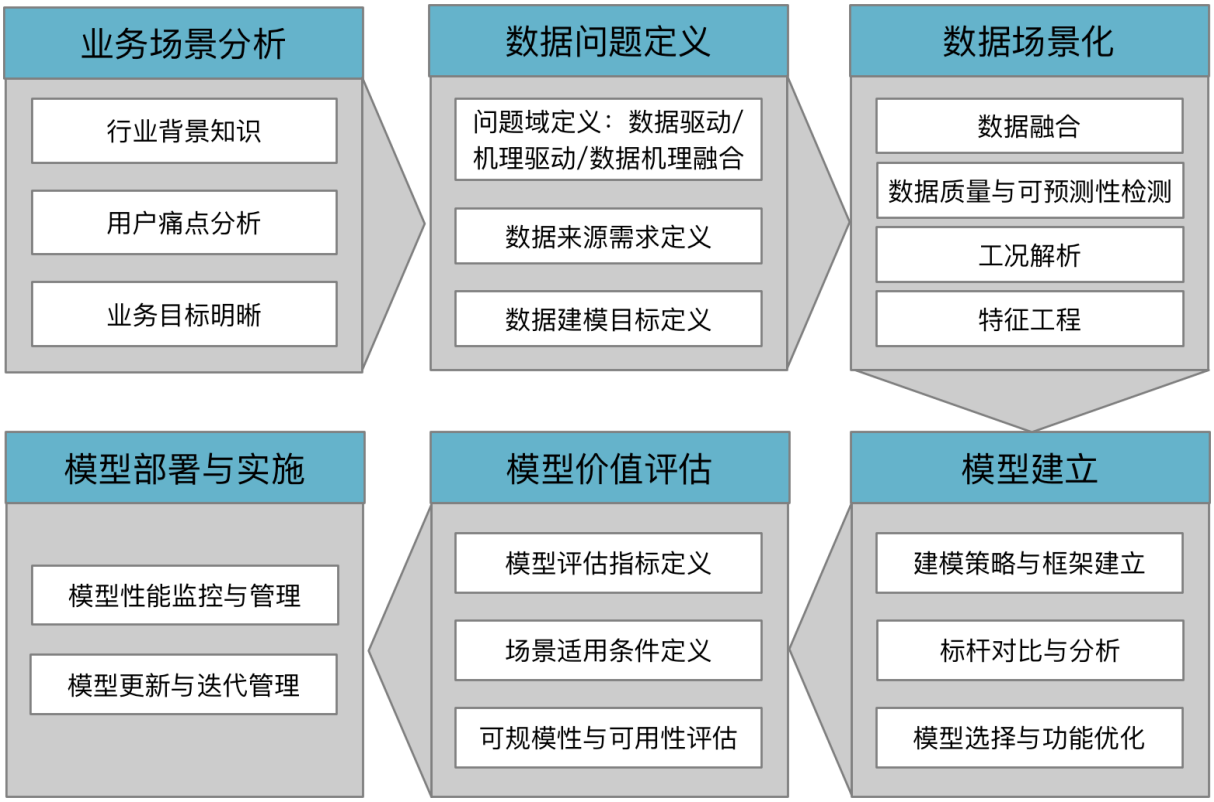


图 2-2：工业大数据的建模过程

下面，我们将针对本次工业大数据竞赛的题目，将其作为一个案例对工业大数据的建模过程展开讨论。

（二）案例-风机结冰故障

1.结冰预测的整体分析思路

大量运行经验表明，风机叶片结冰会改变叶片叶形，破坏叶片气动特性，从而导致风机效率下降和运行不稳定，进而对电网的稳定运行产生影响。因此，实现早期叶片结冰预测可以有效提高风机运行效率和电网运行安全。

风机结冰预测分析由物理建模、特征提取、动态特性分析、建立预测模型和诊断分析几个部分组成。整体分析流程如图 2-3 所示。叶片结冰可以看做一个缓慢的能量累积和转化过程，结冰的程度、影响度与环境条件（温度、湿度、风速等）、以及风机参数（叶型、高度、额定功率等）有关。为了建立准确的结冰预测模型，实现结冰早期和全过程的预测诊断，首先要对结冰的物理过程和风机参数对结冰影响的特性进行分析，充分了解结冰过程中的能量累积、转换和守恒规律，在此基础上提取能够表征结冰程度的关键参数。然后，定量分析叶片结冰状态对风机效率的影响关系，在此基础上提取表征风机受结冰影响的性能参数。因此，物理模型的分析从以下两个角度入手:

- 风机叶片结冰动力学模型
- 叶片结冰状态对风机性能的影响模型

这两种方法能够从两个视角透彻分析结冰过程和风机性能之间的作用机理。一方面，建立叶片结冰动力学模型能够从本质上展示水蒸汽在金属表面的能量转换和累积过程，使我们能

够利用给定的条件和数据定量刻画结冰的严重程度；另一发面，建立叶片结冰状态对风机性能影响的关系模型，使我们能够利用叶片结冰程度指标定量表征风机整体性能，从而实现完全封闭的从观测数据到风机性能的结冰关系模型。利用这个关系，我们可以从中进一步提取相应的特征作为训练模型的条件属性，作为结冰预测模型的输入参数。

在实际运行中，严重的结冰一般能够被轻易检测到，并通过风机除冰系统自动除冰。然而，除冰系统却难以检测早期结冰状态。虽然叶片在结冰早期产生了一定的变形，但对机组的性能影响不明显，因此难以察觉。早期的叶片结冰在不处理的情况下一般都会演化成严重结冰，所以，分析结冰全过程的动态特性对结冰预测来说十分重要，它能够帮助预测模型实现结冰早期的监测和诊断。

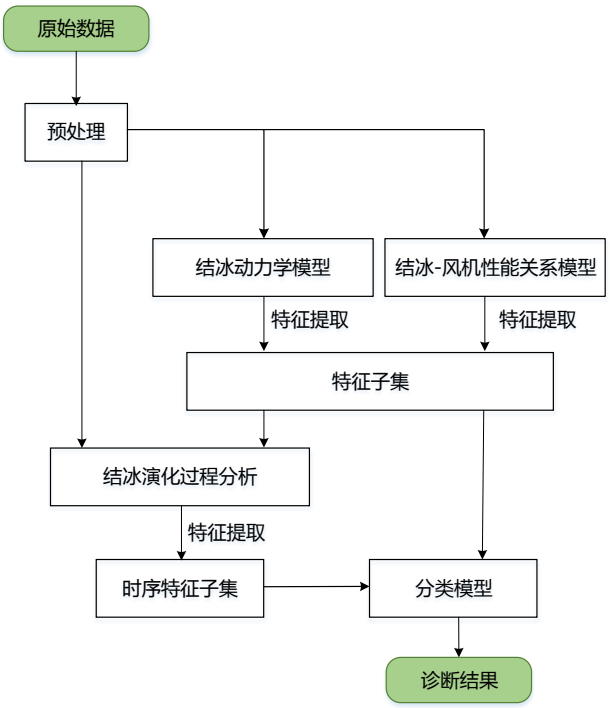


图 2-3：风机叶片结冰预测分析流程

在建立预测分类模型的时候，需要考虑风机结冰数据的类不平衡问题。一般来说，对数据进行重采样能够有效降低类不平衡带来的建模误差。将结冰样本进行过采样，将非结冰样本进行欠采样，或者两者同时进行，以达到结冰和非结冰样本在模型训练时有基本相近的比例。如果结冰样本足够多，也可以选择对类不平衡问题不敏感的分类模型进行建模。

2. 风机结冰物理模型与特征提取

根据 Makkonen 关于物体表面覆冰模型[1]，在给定环境温度的条件下，单位时间内物体表面的覆冰质量可由如下关系表示：

$$dM = \alpha_1 \alpha_2 \alpha_3 \rho V_\infty S \cdot dt \quad (1)$$

其中，$\alpha_1, \alpha_2, \alpha_3$表示结冰状态的 3 个系数，分别为：撞击系

数、黏着系数和增长系数。撞击系数表示水滴从无穷远能够成功撞击物体表面的比例（概率）；黏着系数表示水滴撞击物体表面后能够附着而不反弹的比例（概率）；增长系数表示水滴附着在物体表面后能够持续存在而不融化的比例（概率）。ρ表示空气中水蒸汽密度（含水量），V_∞表示无穷远来流速度（风速），S表示物体表面积，为常数。

在风机结冰预测问题中，由于风机叶片各个截面所处地点几乎相同，叶片大小空间尺度远远小于空气含水量变化的空间尺度，因此含水量ρ可以认为是定值。黏着系数α_2在绝大多数情况下均为 1。因为当温度在-5℃～0℃时，物体表面会产生雨凇结冰。此时物体表面会被一层粘性液体所覆盖，此时水滴几乎不能从撞击中逃脱。当温度低于-5℃时，会产生雾凇结冰，这时由于温度过低水滴在接触物体表面会瞬间凝结成冰，也不会从物体表面逃逸出去。增长系数α_3在雾凇结冰时为 1，因为水滴完全冻结，不会融化。在雨凇结冰时，α_3和含水量有关，一般情况下，含水量越低，α_3越大。由于含水量是定值，因此在风机叶片结冰预测的问题中，增长系数α_3也是常数。因此 Makkonen 覆冰模型可以简化为：

$$dM = C \cdot \alpha_1 V_\infty \cdot dt \quad (2)$$

为了能够提取表征叶片结冰质量的特征参数，我们并不关心常数的具体取值，只关心能够用观测数表征结冰质量的函数形式。因此只要确定由观测参数表征撞击系数的形式，结冰质

量就可以通过观测参数表征。Finstad 等人[2]通过半经验公式拟合了撞击系数α_1:

$$\alpha_1 = A - 0.028 - C(B - 0.0454) \quad (3)$$

其中:

$$A = 1.066K^{-0.00616}\exp\left(-1.103K^{-0.688}\right),$$
$$B = 3.641K^{-0.498}\exp\left(-1.497K^{-0.694}\right),$$
$$C = 0.00637(\phi - 100)^{0.381}.$$
$$K = \frac{\rho d^2}{9\mu D}, \phi = \frac{Re^2}{K}, Re = \frac{\rho_a dv}{\mu}$$

d为水滴直径，在叶片结冰过程的时间尺度内，近似为定值；D为物体截面平均直径；ρ_a为湿空气密度，为定值；v为来流速度，等于V_∞；μ为空气的粘性系数，为常数。

通过上述分析可知，撞击系数和来流速度v具有直接的对应关系，其他参数由于全部是常数，因此可以推断出:

$$\alpha_1 \propto f(v) \quad (4)$$

将（3）（4）代入（2）可以导出结冰质量和风速之间的关系，即:

$$\frac{dM}{dt} \propto V_\infty^{1.762} \quad (5)$$

上式说明，在给定温度条件下，空气含水量、空气密度、水滴形状均不变的情况下，单位时间内叶片结冰质量取决于风速的大小，风速越大，结冰质量越大。当在利用观测数据建模时，经过归一化后的风速，利用近似的风速的平方也能较好的拟合上述模型，因此，在误差允许的范围内，可以用$\frac{dM}{dt} \propto V_\infty^2$模

型代替。

当温度发生变化时，结冰质量的关系模型变为:

$$\frac{dM}{dt} \propto f(T) \cdot V_{\infty}^{2} \quad (6)$$

另一方面，为了能够更加直观的表示结冰质量对风机功率的影响，需要进一步利用功率-结冰质量关系来验证（6）模型。在非结冰情况下，风机会按照正常模式下的风机功率特性曲线工作，发生结冰后，实际运行的功率曲线会偏离正常模式，而这个偏离程度意味着结冰的严重程度，和结冰质量是等价的。

首先，利用模糊自适应神经网络对训练数据的正常样本拟合功率特性曲线得到功率特性曲线的基线模型，然后利用该模型估计测试数据下的功率输出。测试数据包含结冰和非结冰数据，那么结冰严重程度可以表示为:

$$W_{\beta} = \frac{P_{real} - P_{pred}}{P_{real}} \propto \frac{dM}{dt} \propto f(T) \cdot V_{\infty}^{2} \quad (7)$$

P_{real}, P_{pred} 分别表示测试数据的实际功率和通过基线模型估计的功率。（7）表示风功率的相对残差，图 2-4 显示了结冰数据（红色部分）和非结冰数据（蓝色部分）的功率相对基线模型的偏离程度。图 2-5 表示了观测数据的结冰严重程度和风速的关系。

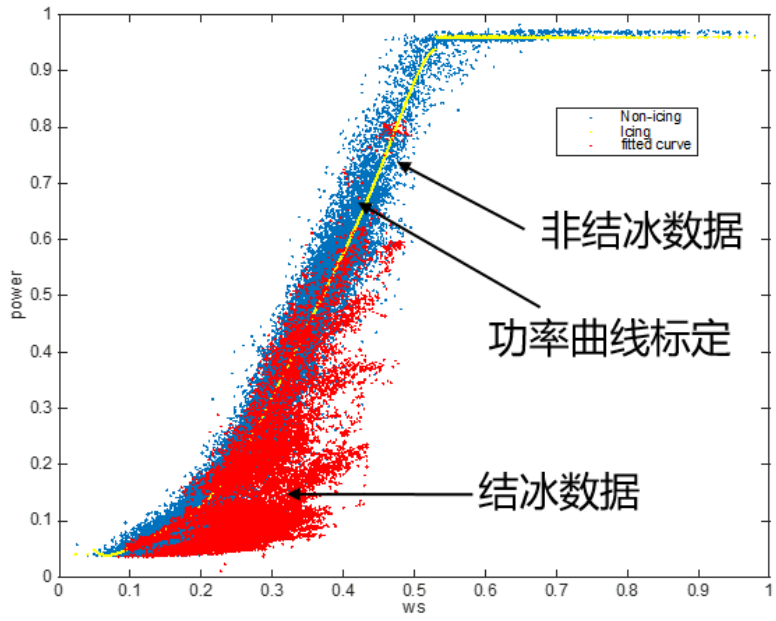


图 2-4：功率特性线拟合

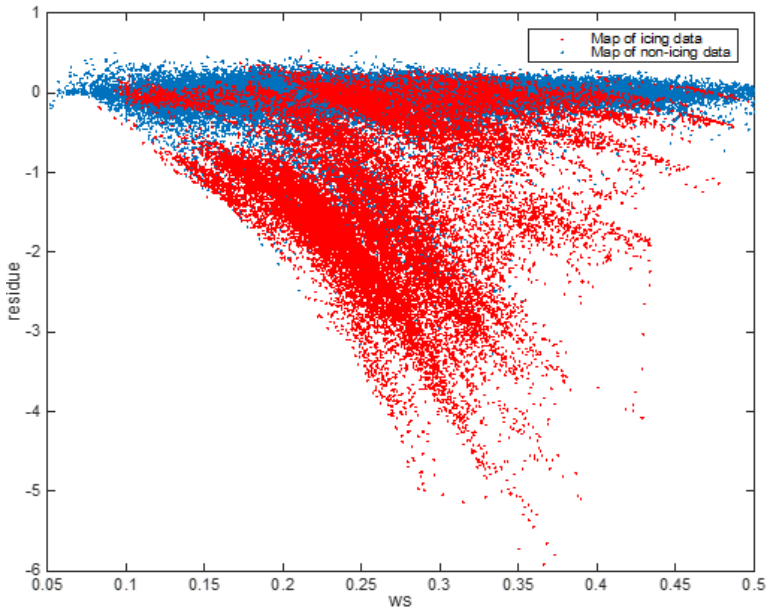


图 2-5：结冰质量—风速关系

通过图 2-5，我们可以非常明显的区分大部分的结冰数据和非结冰数据。在结冰的状态下，结冰质量和风速具有明显幂律关系，虽然这个关系还受到环境温度的调制作用，但是在非结冰数据上，这种关系是不存在的。因此，我们可以提取结冰

预测的第一组重要的特征：结冰质量的度量参数（V_∞^2，T）、结冰严重程度（w_β）。

风机功率对风速的响应模型进一步解释了为什么风功率的偏差能够表示结冰的严重程度。Rahimi 等人[3]揭示了风机功率和风速之间的关系：

$$P=\frac{1}{2}\rho_a\cdot S\cdot C_p\cdot V_\infty^3 \quad (8)$$

这里C_p为风能利用率，S为叶片面积。在正常情况下，$C_p=1, S=\text{constant}$；而在非结冰状态下，风能利用率小于 1，并且叶片面积也会由于覆冰而稍加改变。该模型说明在风速一定的情况下，风机输出功率和结冰程度有明显的单因素对应关系。定义风能综合利用率：$C_{total}=S\cdot C_p$。在图 2 所示的功率特性曲线的基线模型下，可以将实际功率和拟合功率分别用（8）式来表示：

$$P_{real}=\frac{1}{2}\rho_a\cdot S_{real}\cdot C_{real}\cdot V_\infty^3$$

$$P_{pred}=\frac{1}{2}\rho_a\cdot S_{pred}\cdot C_{pred}\cdot V_\infty^3$$

$$S_{pred}=\text{constant}，C_{pred}=1$$

基线模型表征正常情况下的功率-风速响应，因此可以定义在任意时刻的风机风能综合利用率：

$$C_{total}=\frac{S_{real}\cdot C_{real}}{S_{pred}\cdot C_{pred}}=\frac{P_{real}}{P_{pred}}$$

风机风能综合利用率可以作为另一个重要的结冰预测模型的特征，它是结冰严重程度对风机性能影响的量化指标。

通过对风机叶片结冰过程的机理分析和物理建模，我们提取了若干表征结冰的状态和本质属性的特征，如表 2-1 所示这些特征能够作为建立预测模型的输入参数，能够最大程度地为分类算法提供相互独立的信息，在保证预测精度的前提下，降低分类模型的复杂度，提高模型的泛化性能。

表 2-1：风机叶片结冰预测模型瞬态特征

特征	描述	单位
W_β	结冰严重程度	—
V_∞	平均风速	m/s
V_∞^2	风速平方	m^2/s^2
P_{real}	输出功率	kW
C_{total}	风能综合利用率	—
WindDirection	风向角	—
ρ	相对湿度	%
T	平均环境温度	℃

3.风机结冰过程演化分析

风机叶片结冰是一个缓慢的能量累积过程，在结冰早期由于现象不明显难以发现，然而早期结冰的检测对于机组健康运行至关重要，如果能够检测出早期的结冰状态，则能够防止机组由于严重结冰导致的经济性和安全性下降的问题。然而，仅通过上述物理模型提取的瞬态特征难以实现准确的早期结冰预测，需要通过对结冰过程的演化规律进行分析，提取结冰演化过程中的统计特征和时序特征，才能更好的实现早期结冰预测。

图 2-6 为在结冰质量-风速的关系中，早期结冰数据（橙色

圆圈内的红色数据）的状态。可以看出，这些早期结冰数据的结冰严重程度并不高，和非结冰数据（蓝色部分）几乎重合，这部分数据通过观测数据和瞬态特征不能很好的分类，因此需要提取其他特征来解决这一问题。

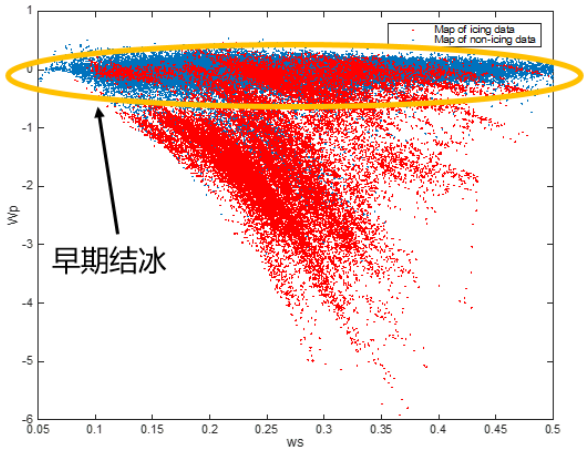


图 2-6：早期结冰数据

风机结冰的过程具有很强的周期性，结冰周期一般为早期结冰-明显结冰-严重结冰-除冰-不结冰。因此为了分析结冰全周期的演化规律，首先需要将原始数据序列化，将其分割成若干片段。每个片段为一个完整的结冰过程的数据，或者一段给定的连续时间的非结冰数据。通过训练数据的标签和除冰设备开启时设备振动的周期性变化，将原始数据分割成时间序列片段。图 2-7 和图 2-8 分别为一个结冰周期和一个非结冰片段在结冰质量-风速关系中的分布模式。在一个结冰周期内（1-2 小时），环境温度的变化可以忽略不计，从图 2-7 红色部分可以看出，风速和结冰质量近似满足 $\frac{dM}{dt} \propto V_{\infty}^{2}$ 的关系，而曲线的统计特性能够定量展示了一个结冰周期内风速、结冰程度的综合演

化规律。图 2-8 的绿色部分表示一个给定时间段（2 小时）内非结冰的数据分布，显然该数据并不满足覆冰模型的关系。

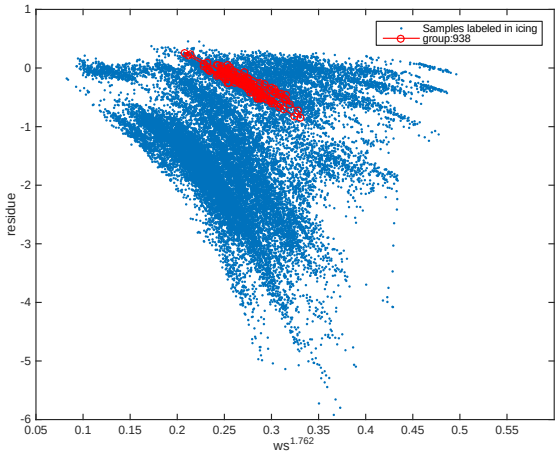


图 2-7：一个周期内结冰数据的分布

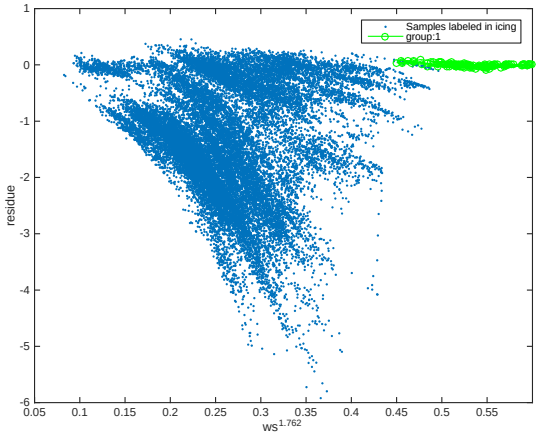


图 2-8：一个时间周期内非结冰数据的分布

如果我们能从序列数据中提取出易于分类的特征，就能够准确地检测早期结冰。因此，在设计预测模型时，不仅要提供瞬态特征，还要利用滑动窗提取给定长度下的统计特征。这些统计特征能够很好地反映这一段序列数据的演化规律和状态，因此能够更好的发现早期结冰。为了从瞬态特征中提取演化规律信息，我们将表 2-1 的特征利用滑动窗进行进一步处理，得

到一系列统计特征，如表 2-2 所示。

表 2-2：风机叶片结冰预测中的统计特征

特征	描述	单位
$W_{\beta}max$	结冰严重程度最大值	—
$W_{\beta}acc$	结冰严重程度累积	—
$W_{\beta}std$	结冰严重程度标准差	—
$W_{\beta}slope$	结冰严重程度平均变化率	—
$C_{total}max$	风能综合利用率最大值	—
$C_{total}acc$	风能综合利用率下降累积	—
$C_{total}std$	风能综合利用率标准差	—
$C_{total}slope$	风能综合利用率平均变化率	—

4. 风机结冰预测分类模型

通过风机叶片结冰机理建模和演化规律分析，我们从原始数据中提取 16 个特征用于建立基于数据驱动的结冰预测分类器。本文的原始数据来源于《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。在建立分类模型前需要对原始数据进行预处理，包括:

- 剔除离群样本和坏点;
- 核对样本标签;
- 原始数据等间隔重采样（1s 间隔）;

- 数据归一化，将主要参数按照训练集标准化到均值=0，方差=1 分布;
- 原始数据按照滑动窗（L=2h）片段化，得到序列数据;
- 设置训练集和测试集。设置训练集为第一个月数据，测试集为第二个月数据。
- 训练集消除样本类不平衡;

消除类不平衡的方法一般采用大类样本欠采样、小类样本过采样、改变两类样本权重等方法。本文采取的方法是大类样本欠采样，最后训练集样本容量约为 4 万，正负类样本比例约 1:1。

在训练集上提取上述特征，得到完整训练集。构建支持向量回归模型作为基分类器，核函数为径向基函数。混淆矩阵作为分类模型的评价指标。

TP 表示为所有非结冰样本中分类为非结冰样本的比例;

TN 表示为所有结冰的样本中分类为结冰样本的比例;

FP 表示为所有结冰的样本中分类为非结冰样本的比例;

FN 表示为所有非结冰样本中分类为结冰样本的比例。

表 2-3 给出了测试集上结冰和非结冰的预测效果。在原始数据质量不佳和采用单个分类模型的前提下，非结冰和结冰数据均能够达到很高的分类正确率。分类结果也印证了特征对分类模型效果的重要性。

表 2-3：叶片结冰预测效果

	预测非结冰	预测结冰
实际非结冰	0.94（TP）	0.06（FN）
实际结冰	0.18（FP）	0.82（TN）

5.总结

风机叶片结冰预测是典型的健康预诊模型的应用。叶片结冰过程的机理复杂，早期结冰又难以发现，并对设备的经济安全运行产生很大的影响。叶片结冰预测关键在于如何能够在结冰形成过程中尽可能早地监测到结冰的风险，通过结合结冰机理和数据分析的方法能够提取与结冰状态趋势相关的特征，从而提早预知结冰概率，提高预诊断的准确度与及时性。因此，一方面通过建立叶片结冰的物理模型，分析叶片结冰质量和风速之间的对应关系，寻找叶片结冰质量对机组功率的影响，提取表征叶片结冰状态的瞬态特征；另一方面利用覆冰模型研究结冰过程的演化规律，从瞬态特征中提取表征结冰过程的统计特征，从而实现结冰早期诊断。实验结果表明，通过该方法建立的结冰预测模型能够在最简单的模型结构下获得最佳的分类效果。

6.参考文献

[1] Makkonen L. Models for the Growth of Rime, Glaze, Icicles and Wet Snow on Structures[J]. Philosophical Transactions Mathematical Physical & Engineering Sciences, 2000, 358(1776):2913-2939.

[2] Finstad K J, Lozowski E P, Gates E M. A Computational Investigation of Water Droplet Trajectories[J]. Journal of Atmospheric & Oceanic Technology, 1958, 5(1):160-170.

[3] Rahimi E, Rabiee A, Aghaei J, et al. On the management of wind power intermittency[J]. Renewable & Sustainable Energy Reviews, 2013, 28(28):643-653.

三、首届工业大数据创新竞赛解法集

针对这次结冰数据竞赛题目，本白皮书收录的解法如下:

#	方法	所属单位
1	基于 CNN-LSTM 深度学习网络的风机叶片结冰预测	万腾科技
2	基于物理原理+KNN 分类的混合预测模型	济中节能
3	基于领域知识特征构建和未来结冰概率估计的风机叶片结冰预测	浙江运达风电股份有限公司
4	基于数据驱动和非均衡数据学习的故障预测研究	北京邮电大学
5	基于敏感特征的风机叶片结冰预测算法	西安交通大学

下几节将对每个解法从数据解析、方法、以及验证步骤等方面展开详细的讨论。

（一）基于 CNN-LSTM 深度学习网络的风机叶片结冰预测

1.团队介绍

团队名称: 万腾科技

成员姓名	团队角色	职位
张嗣昌	队长	算法部负责人兼副总裁
侯振寰	队员	高级算法工程师
侯宗波	队员	高级算法工程师

万腾科技参赛队是以山东万腾软件部为基础，本着实践探索创新的研究精神参加本届大赛。张嗣昌担任队长，负责模型选型，建模与优化。张嗣昌毕业于山东大学，基础数学硕士，现任万腾科技算法部负责人兼副总裁，主要研究最优化算法，人工智能，模式识别三维重建。主持研发了高级计划排产系统，设备故障预测与健康管理系统，工业仿真及工业 AR/VR 等项目。侯振寰负责建模与优化。她是天津大学电路与系统专业硕士，

现任万腾科技高级算法工程师，研究方向为 Deep Learning，在 SCI、EI、中文核心期刊上发表论文三篇。参与设备故障预测与健康管理系统的研发。侯宗波负责特征工程与模型选型。他毕业于山东中医药大学，生物医学工程硕士，现任万腾科技高级算法工程师，研究方向为最优化算法，人工神经网络。参与了高级计划排产系统，设备故障预测与健康管理系统的研发。

2.背景简介与文献调研

风机叶片结冰是一个缓慢推进的过程，与外界环境因素息息相关，如环境温度、风速、空气湿度等等，成因非常复杂，是风电领域的一个全球范围难题。叶片结冰会导致叶片加重，容易造成叶片折断，存在很大的安全隐患，此外，叶片结冰会影响风力发电工作效率，气候比较恶劣的地区年发电效率会减小 20%至 50%[1]。因此，关于叶片结冰的研究具有非常重要的现实意义。

近些年，国内外在风机叶片结冰问题上取得许多可喜成果，研究主要分为三个方面：风洞实验研究、数值模拟研究、防冰除冰研究[2]，但是针对风机叶片结冰预测的研究还比较少。目前针对叶片结冰故障的监测手段主要是比较风机实际功率与理论功率之间的偏差，当偏差达到一定值后会触发风机的报警和停机。然而，此时叶片已经大面的积结冰，叶片折断风险大大增加。此外，虽然许多新型风机都设计了自动除冰系统，但是很难预测到结冰初期。

本文在深度学习的方向上对风机结冰预测进行探索，根据

提供的数据训练模型，通过模型预测叶片是否结冰，以便尽早进行除冰处理。预测一般可以划分为分类或回归问题，根据评分规则，我们将本问题划为分类问题，此外，竞赛提供的数据是一系列连续时间上的测量值，属于时间序列预测问题，因此可以尝试一些时序预测算法。最终我们选用了卷积神经网络与长短期记忆网络组成的深度学习模型，验证结果表明预测精度优于其他传统的机器学习算法。

3.数据解析

竞赛共提供了 5 个不同风机的 SCADA 采集的数据集：训练集 15、21，初赛测试数据集 08，复赛测试数据集 10、14。训练数据与初赛测试数据集包含 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。复赛测试数据集中删除了 group 维度。由于测试集数据中时间维度量化为以 1 开始的等差序列，掩盖了实际的物理意义，因此时间维度同样不能用于模型训练。在实际的模型训练中时间维度可能是非常有用的特征，最直观的是与环境温度有联系，比如在我国内蒙古地区，昼夜温度相差非常大，夜晚较于白日更易结冰。综上，最终剩余数据集中的 26 个变量可用于进一步特征选择。

表 3-1 为训练集 15、21 的基本情况，其中训练集 15、21 的故障数据所占数据总量的比例分别为 16.5:1、17.9:1，正负样本非常不平衡。

训练集	数据量	故障数据	维度	时间跨度

风机 15	393886	23892	28	2 个月
风机 21	190494	10638	28	1 个月

表 3-2 为三组测试数据的数据量以及维度说明，赛委会将复赛的测试数据进行了随机删除，我们粗略地将风机 10、14 的数据量与风机 08 进行对比，数据量较之减少 3 至 4 万条，可见数据缺失情况还是比较严重的，在之后的内容中还会用更合理的方式对这一情况进行说明。

表 3-1：训练集数据统计

训练集	数据量	故障数据	维度	时间跨度
风机 15	393886	23892	28	2 个月
风机 21	190494	10638	28	1 个月

表 3-2：测试集数据统计

测试集	数据量	维度
风机 08	202328	28
风机 10	174301	27
风机 14	163732	27

此外，训练数据中两条数据之间的时间间隔多为 7 秒、8 秒、10 秒，但是由于存在设备停机、删除数据等情况，记录的数据会存在间断，即时间间隔比较大。这些间断会给预测带来干扰。下图为训练样本中抽取的一处数据不连续的例子。

2015/11/1 20:59	1.373182	1.250185	1.59864	0.199146
2015/11/1 20:59	1.800986	1.336604	2.223423	0.176703
2015/11/2 10:07	0.606085	1.26348	0.5994	-0.2031
2015/11/2 10:07	0.646652	1.273452	0.61543	-0.2756

图 3-1：数据不连续样例

经过对数据的一系列分析，影响模型预测性能主要可以归纳为以下四个方面:

- 样本中存在停机、人为删除数据、无效数据等，会造成某些时间段的数据缺失
- 选取哪些特征用于模型预测
- 何种训练模型比较适合本问题
- 叶片正常与结冰故障数据不均衡

4. 方法

方法分析流程:

(1) 数据预处理

训练数据集中包含正常数据、故障数据以及无效数据，分别将其对应的训练样本标签 label 标记为 0、1、-1。其中，正常数据、故障数据、无效数据所占比例约为 16.8:1.1:1，三种数据分布非常不均衡。对于无效数据，我们无法将其划分成正常数据或是故障数据，因此在实际训练中将其删除。此外，为了提高模型的鲁棒性，我们根据训练数据中的 group 维度随机地删除部分非结冰数据以及结冰严重数据，经过剔除数据处理，正负样本所占比依然存在很大的差异，这种数据不平衡会对模型最终的叶片结冰预测有非常大的影响，一般处理方式有三种:

欠采样、过采样、在模型 loss 函数中增加惩罚项以及模型集成[3]。由于我们想尽可能的保留原始数据，因此选择第三种方法，并且选取了更加适合此种情况的评价函数。

在之前的内容中，我们提到数据不连续的问题，连续时间内传感器采集的数据趋势相对一致，但是有时间间隔的两段数据会存在差异，因此，有必要将数据分割为多个连续子时间段。数据不连续的位置，其属性前后时刻差分值应该能有所表征，经过试验，我们最终选取 pitch1_ng5_tmp、pitch2_ng5_tmp、pitch3_ng5_tmp 这三个属性，其前后时刻数值差分比较大的位置用于连续子序列的分割。在下方表 3-3 中，记录了我们将数据分割成连续数据段的段数，从中可以看出测试集 10、14 子时间段个数超过 300，说明了这两个数据集数据缺失严重。

表 3-3: 数据分割子时间段统计

训练集	子时间段		测试集	子时间段
	原始数据	按 group 删除后	风机 08	76
风机 15	173	640	风机 14	310
风机 21	90	357	风机 10	302

(2)特征加工与选择

接下来我们将可用的 26 个特征进行加工、选择。26 个特征两两绘图，其中同一风机 3 个叶片的叶片角度、速度、变桨电机温度数据分布较为一致，如图 3-2 所示，求其均值作为新

特征记作 mean_pitch_angle、mean_pitch_speed、mean_moto_tmp。

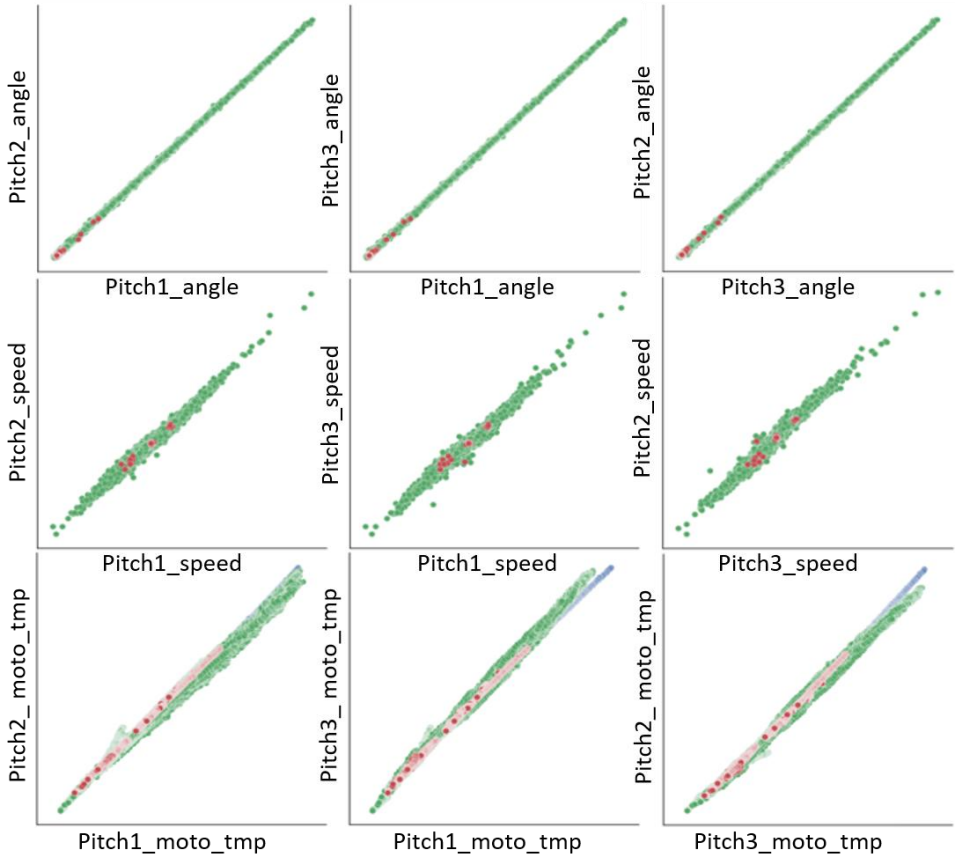


图 3-2：叶片角度、速度、变桨电机温度数据分布

在数据进行分割之前，计算 pitch_angle、pitch_moto_tmp、pitch_ng5_tmp 这三组特征的前后时刻差分值，得到 diff_pitch_angle（叶片角度前后时刻差分）、diff_pitch_tmp（变桨温度前后时刻差分）、diff_ng5_tmp（ng5 温度前后时刻差分），这三个特征能够在一定程度上表征数据间断。

我们对现有的特征进行组合，其中，风速与网侧有功功率的比值、风速与发动机转速的比值以及风速与网侧有功功率、

发动机转速之间的比值，在叶片结冰期间会有明显的上升趋势，能够很好的表征叶片结冰过程。根据实际数据进行建模生成三个特征：r_windspeed_to_power、r_windspeed_to_generato r_speed、r_square，其中，wind_speed、generator_speed、power 分别为风速、网侧有功功率、发动机转速的特征，建模的公式如下。图 3-3 为从训练样本中抽取一段时间的三个特征的波形图，在叶片结冰的区域内（label 为 1 的区域内），三个特征在中间部分开始一直递增，直到结冰过程结束。

$$r\_windspeed\_to\_power = (\frac{wind\_speed + 5}{power + 5})^2 - 1 \quad (1)$$

$$r\_windspeed\_to\_generator\_speed = (\frac{wind\_speed + 5}{generator\_speed + 5})^2 - 1 \quad (2)$$

$$r\_square = [\frac{(wind\_speed + 5)^2}{(power + 5) \times (generator + 5)}]^2 - 1 \quad (3)$$

图 3-3：特征组合波形

另外，对特征 r_square 采用不同参数分段进行最小二乘多项式拟合，可以得到拟合曲线的斜率，在这里分别选取参数 64、128 的拟合斜率 slope_fit1、slope_fit2 作为特征。

(3) 算法选择

深度学习能够自学习特征，在某种程度上可以弥补传统机器学习对于特征要求高的缺点[4]。因此，本文建立 CNN+LSTM 二分类深度学习网络用于叶片结冰预测，并且采用 Keras 深度学习框架进行网络的搭建。Keras 非常适合模型的快速搭建，同时也提供了很多接口供使用者自定义[5]。

模型设置为七层，采用顺序结构排列，由两层卷积层、两层 LSTM、三层全连接层组成。其结构如下图 3-4。

图 3-4：训练模型结构

模型的主要工作原理是：训练数据首先经过两层卷积层，初步地学习多个连续时刻数据的特征以及变化趋势，接着使用长短期记忆网络 LSTM 进一步学习不同时间段的特征，最终达到叶片结冰故障预测的目的。

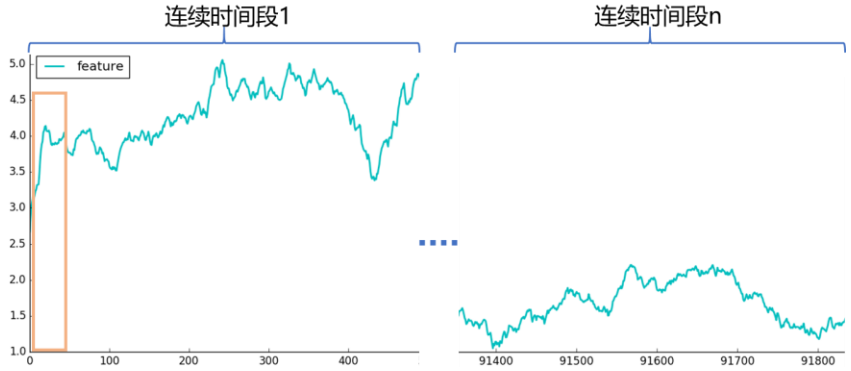


图 3-5：滑窗数据示意图

对于模型的输入数据，我们进行了滑窗选取。根据数据分割处理，训练数据被分割成 n 个连续子时间段，分别对连续的子区间内，采用固定的步长滑动选取一定窗宽的数据，这里我们选取的是窗宽为 64、步长为 1。窗宽设置比较大能够包含更多连续时刻信息，但是数据量会激增，但是选取过小，所包含的时间维度信息又可能不够。通过滑窗处理，我们的输入数据将同时包含时间维度的信息以及特征维度的信息。

由于，本赛题提供的训练、测试数据中结冰数据少于正常数据（在风机 15、21 的训练数据集中，叶片正常数据与结冰数据比值约为 16:1），我们选用了马修斯相关系数（Matthews Correlation Coefficient, MCC）作为评价模型分类的准则。MCC 广泛应用于评价机器学习中的分类问题，尤其是正负样本不均衡情况下的分类问题[6]。

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

其中，TP 为正样本预测正确个数，TN 为负样本预测正确个数，FP 为正样本预测错误个数，FN 为负样本预测错误个数。MCC 求解的是二进制分类中真实值和预测值之间的相关系数，其取值范围为[-1,1]，数值越接近 1 代表模型预测的越准。

同样，Loss 函数中增加有关正负样本不平衡的惩罚项，既可以提高训练模型的准确度，又能在一定程度上减轻数据不平衡带来的影响。在 Loss 函数中引入有关马修斯相关系数的惩罚项，其公式如下:

$$Loss = binary\_crossentropy + \lambda * (1 - mcc) \quad (5)$$

其中，$binary_crossentropy$二进制交叉熵损失函数；λ为惩罚因子；mcc为当前模型的马修斯相关系数。

此外，模型中引入了 Dropout 以及正则化来减小模型过拟合的概率，Adam 优化器的学习率成指数下降，在一定程度上避免了由于学习率引起的，随着迭代次数增大 Loss 没有减小的情况，防止模型欠拟合。

模型的工作流程框图如图 3-6，根据在线评测的测试集在模型中的预测表现情况，进一步回调模型参数并且进行特征选择，删除对预测作用不大的特征，减少数据维度，降低训练复杂度，从而优化模型的预测性能。

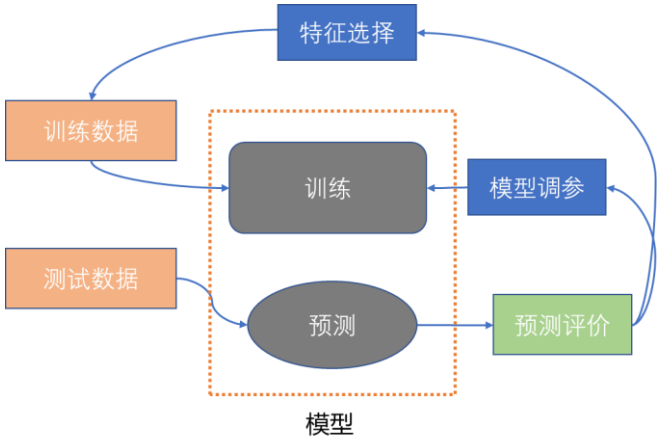


图 3-6：工作流程

(4)方法独到之处

1）数据处理方面

a）根据 pitch_ng5_tmp 属性，将训练数据分割成多个连续子时间段；

b）训练数据滑窗处理，丰富了输入数据的特征维度。

2）特征方面

a）根据实际问题，提取了风速与网侧有功功率、发动机转速之间的比值等特征。

3）模型方面

a）组成的 CNN+LSTM 二分类网络用于叶片结冰预测;

b）采用 Keras 深度学习框架快速搭建网络。

4）数据不平衡

a）引入马修斯相关系数用于评价模型预测效果;

b）在 Loss 函数中增加惩罚项。

5.验证

为了涵盖更多的叶片结冰情况，将风机 15、21 的训练数据合并用于模型训练，并且随机抽取 70%做训练集，30%做验证集。模型在训练集、验证集上的马修斯相关系数分别为 0.84、0.68，结果相差比较大的可能原因是随机抽划分训练集、验证集样本，其正负样本的比例存在差异。测试集 08 的预测结果干扰比较少，测试集 10、14 干扰比较多，数据的缺失影响还是未能完全规避。此外，我们还尝试过 SVM、随机森林、Xgboost 等算法训练模型，对于初赛的 08 测试集预测效果也是很不错的，但是预测复赛的两个测试集，其结果较于深度学习的这个模型要差很多。

6.结果分析与经验总结

从初赛开始到复赛最后一次提交结果，在这近三个月可以说收获颇多。之前从未接触过发电风机这一领域，通过不断地查找资料，对问题有了初步地理解。尽管缺少相关领域的专业

知识，但是我们可以从提供的数据出发，利用一些数学方法，分析数据之间以及数据与叶片结冰之间的联系，通过了解数据，可以从中发现一些问题，比如样本不均衡等。根据数据分析结果，我们可以尝试训练不同的模型，选择其中比较合适的模型进一步调试。模型调参是个比较辛苦的过程，可以参考其他人分享的经验。此外，在调整过程中做好记录，把握好调整方向，否则可能在做无用功。

跑程序使用的是 Dell 笔记本 inspiron-7460 内存 8G，从数据处理到模型训练的训练完成大概需要六小时。训练深度学习模型的速度通常比较慢，我们也在预测精度和计算速度上进行了考虑，比如，尽可能的减少冗余输入、优化模型结构等。

本文提出的模型还有很大的提升空间，通过进一步地调整还能达到更好的预测效果；此外，不同风机的数据分布是不同的，这对训练模型的泛化性就有很高的要求，此次复赛的两个测试集由于人为删除大量数据，造成数据分布与训练集、初赛测试集数据有很大差异，可能导致最初训练的模型不再适用，所以，在模型中引入迁移学习能够更好地提高模型泛化性。

7. 参考文献

[1] 王聪, 黄洁亭, 张勇,等. 风电机组叶片结冰研究现状与进展[J]. 电力建设, 2014, 35(2):70-75.

[2] 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.

[3] Japkowicz N, Stephen S. The class imbalance problem: A

systematic study[J]. Intelligent data analysis, 2002, 6(5): 429-449.

[4] Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015: 4580-4584.

[5] http://keras-cn.readthedocs.io/en/latest/

[6] Alexandridis A, Chondrodima E, Paivana G, et al. Music genre classification using radial basis function networks and particle swarm optimization[C]//Computer Science and Electronic Engineering Conference (CEEC), 2014 6th. IEEE, 2014: 35-40.

（二）基于物理原理+KNN 分类的混合预测模型

1.团队介绍

团队名称: 济中节能

成员姓名	团队角色	职位
宋哲	技术顾问	首席科学家
纪浩然	组长	中级工业大数据数据分析师
王成金	组员	中级工业大数据数据分析师
张薇	组员	初级工业大数据数据分析师

济中由美国爱荷华大学能源管理领域专家团队创立，专注于提供整体能源管理与能源数据分析服务。济中重视研发团队打造，团队核心成员曾主持完成《风电预测、并网调度与规划的决策优化模型》国家自然科学基金项目，美国爱荷华州自然资源部能源中心资助的“Data-Driven Performance Optimization of Wind Farms”等能源系统管理和决策优化项目。

2.背景简介与文献调研

目前国内关于叶片结冰故障的诊断仍然处于研究发展阶段，一般都是结冰状态比较严重后进行停机除冰，用于结冰探测的传感器也在发展之中，尚未普及。当前，对于风机叶片结冰故障的诊断主要技术手段是比较风机实际功率与理论功率之间的偏差，当偏差达到一定值后触发风机的报警，但该手段的缺点是当触发报警时，往往是已经发生了叶片大面积结冰现象，即不能在结冰的早期就及时诊断出来。应用大数据和数据挖掘技

术，将是解决风机叶片结冰故障诊断这一难题的有利手段。

3. 数据解析

(1) 数据存在噪音。

a) 数据需要采取去噪音手段，处理异常值、缺失值。

(2) 数据被标准化处理过。

(3) 结冰状态间断。

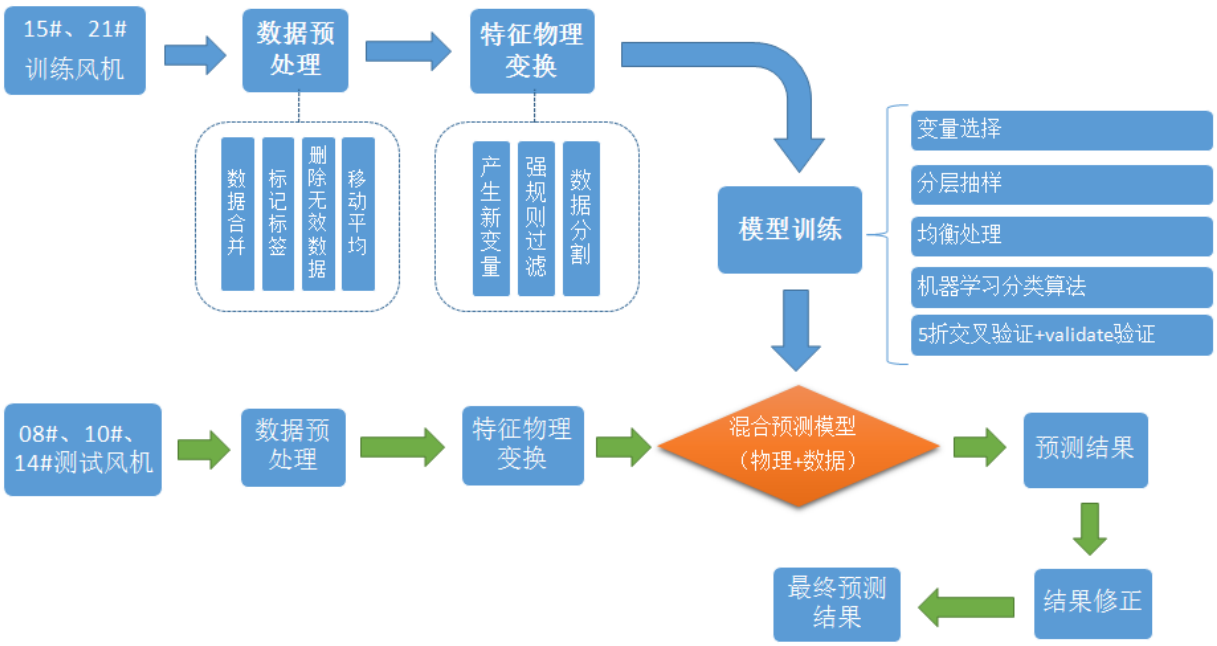
a) 结冰的状态特征表现为每隔一段时间就会持续结冰。

(4) 类别不均衡。

a) 结冰与不结冰数据样本比例失衡，建模过程中需要均衡处理。

4. 方法

■ 设计流程图:

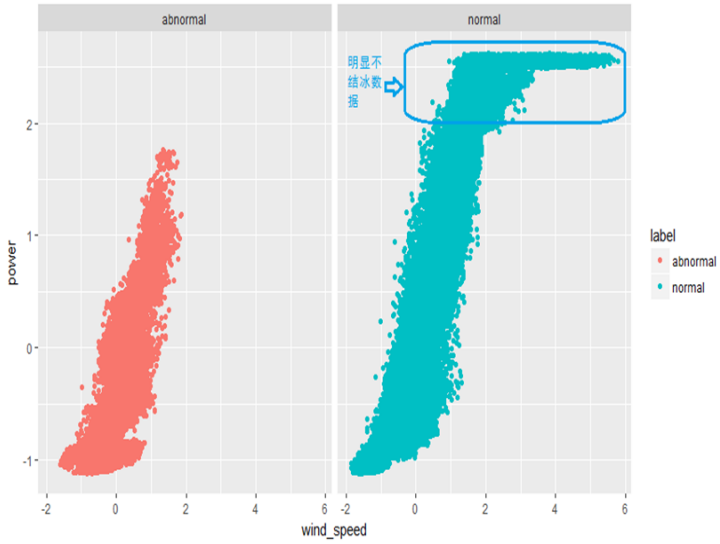


■ 新增变量:

Id	变量名	解释	计算方式
1	Tmp_diff	温差	int_tmp-environment_tmp
2	Torque	扭矩	Power/generator_speed
3	Cp	功率系数	Power/wind_speed^3
4	Ct	推力系数	Torque/wind_speed^2
5	Lambda	速率比	Generator_speed/wind_speed
6	Pitch_angle_Ave	叶片角均值	Average(pitch_angle1,pitch_angle2,pitch_angle3)
7	Pitch_angle_Sd	叶片角标准差	Sd(pitch_angle1,pitch_angle2,pitch_angle3)
8	……	其他变量	

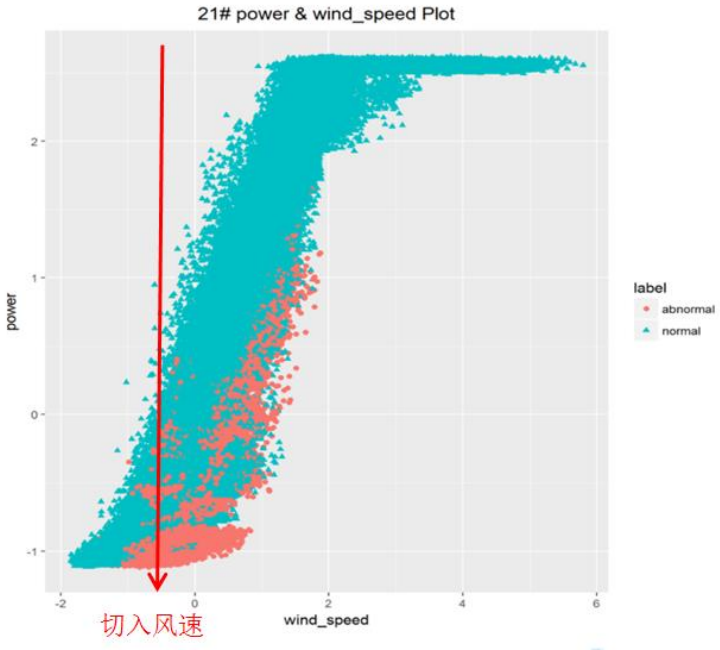
■ 强规则过滤:

通过强规则过滤掉明显不结冰的数据，避免模型过度学习正常数据，提高模型的准确率和泛化能力。模型最终有四条强规则，下图展示其一。



■ 数据分割:

按照切入风速把训练数据切割成两部分，对这两部分数据分别建模，把复杂的模型变成两个简单的模型，避免了单一模型的过度复杂，从而提高了模型的泛化能力。



■ 算法选择:

预测结果较好的机器学习分类算法有 KNN、SVM、ANN，下

表比较了三种分类算法的预测得分结果。此外，C5.0 决策树，randomForest 等分类算法结果并不理想。

算法	交叉验证	validate 验证	Train_15	Train_21	Test_08
KNN	97.15	96.95	92.34	95.62	87.67
SVM	97.26	97.13	88.12	91.54	84.78
ANN	96.83	95.78	90.52	92.21	82.68

综合选择，KNN 算法最优，主要优点有:

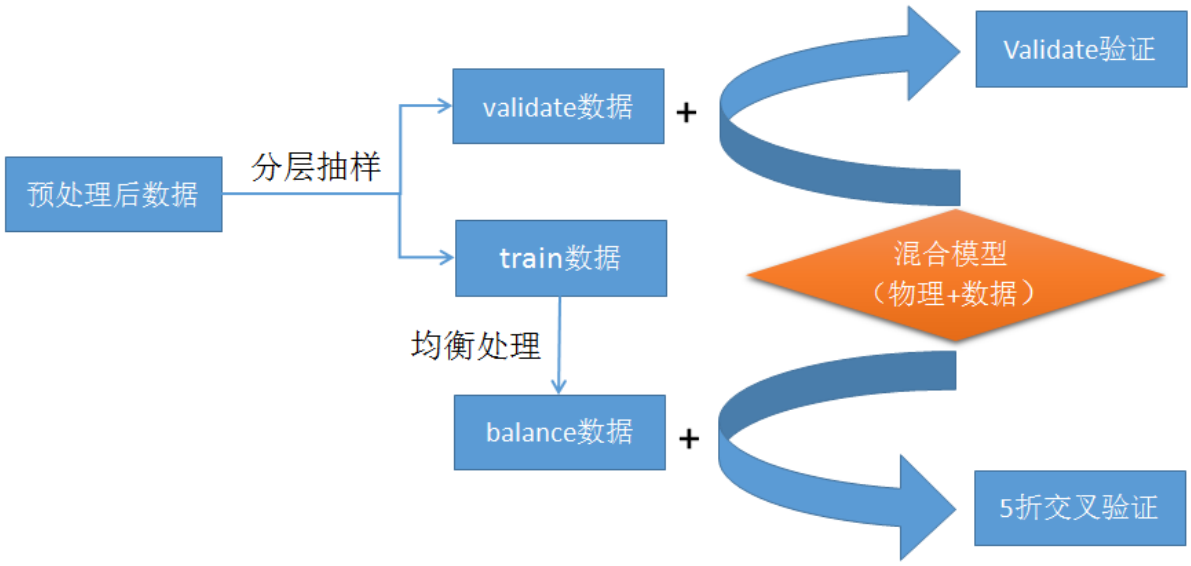
1、算法最简单

2、泛化能力最强

3、模型训练时间更短

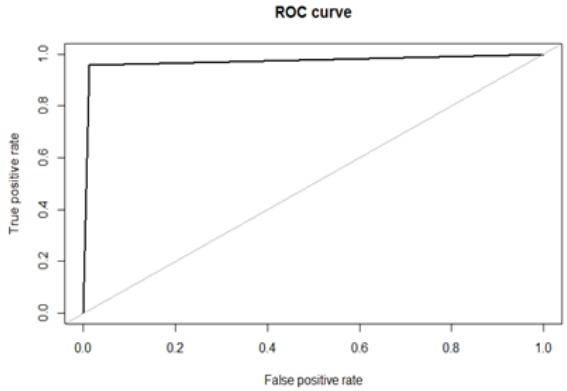
5 验证

- 模型验证流程图:



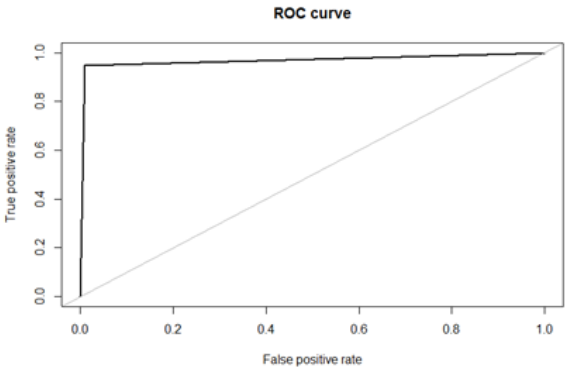
- 模型验证误差曲线:

Validate验证误差曲线



Area under the curve(AUC):**0.972**

交叉验证误差曲线



Area under the curve(AUC):**0.969**

6 结果分析与经验总结

■ 算法结果:

次数	交叉验证	validate验证	Train_15	Train_21	Test_08	final_10、14	排名
第一次	98.45	97.93	91.37	94.72	88.31	81.27	3
第二次	98.69	98.23	92.67	95.26	89.23	**82.01**	**2**

■ 开发环境

1、编程语言: R 语言, 版本: 64bit-3.3.2, 代码行数: 448。

2、编辑器: RStudio。

3、开发环境: 64bit-Win10、CPU i7-2600 3.4GHz、RAM 8G。

4、算法耗时: 555.36 秒。

数据预处理	特征物理变	模型训练	新数据预测	总计

	换			
160.34 s	384.72 s	9.06 s	2.24 s	**555.36 s**

- 算法改进及建议

风机叶片结冰预测算法的一个主要的模型要求是泛化能力，即能够在不同的风机上模型预测效果依然良好，虽然从各个风机的测试结果来看，算法达到了较高的泛化能力（至少 82%的准确率），但算法依然存在可以改进的空间，比如 KNN 参数 k 的选择，强规则的继续排列组合以及分组建模的分割点等。

7. 参考文献

[1] Z. Zhang, Zhe Song and J. Xu, 2015, Data-Driven Wind Turbine Power Generation Performance Monitoring, IEEE Transactions on Industrial Electronics, Vol. 62。

[2] Yu Jiang, Zhe Song, Andrew Kusiak, 2013, Very short-term wind speed forecasting with Bayesian structural break model, Renewable Energy, Vol. 50。

[3] Geng Xiulin.The Elements of Data,Modeling and Decision-making.Beijing: China Renmin University Press,2013。

[4] Winston Chang.R Grapbics Cookbook.Beijing:Posts & Telecom Press,2014。

（三）基于领域知识特征构建和未来结冰概率估计的风机叶片结冰预测

1 团队介绍

团队名称：世属三

成员姓名	团队角色	职位
周书锋	队长	数据应用工程师、CDA 认证业务数据分析师/数据挖掘建模分析师
朱博文	队员	软件工程师、CDA 认证大数据分析师
冯文婷	队员	应用开发工程师、CDA 认证数据挖掘建模分析师

团队来自浙江运达风电股份有限公司，运达风电公司成立于 2001 年，主营大型风力发电机组的设计、生产和销售以及风电场的运行维护、备品备件的供应；并提供风力发电工程的风场规划、技术咨询、设计、施工等服务。队长周书锋曾从事风电叶片结构设计、风电整机载荷计算工作。现为故障预警、性能提升等数据应用提供算法支持；队员朱博文从事风电行业信息化软件研发、大数据分布式平台建设及数据分析相关工作；队员冯文婷从事风电 SCADA 系统开发、风电数据分析应用开发工作。可以看出，“叶片结冰预测”与团队经历中的风电叶片结构设计、载荷计算、大数据平台建设、SCADA 系统开发等多个领域都有密切联系。从领域知识和机理分析方面，我们团队是具有天然的优势的。

2.背景简介与文献调研

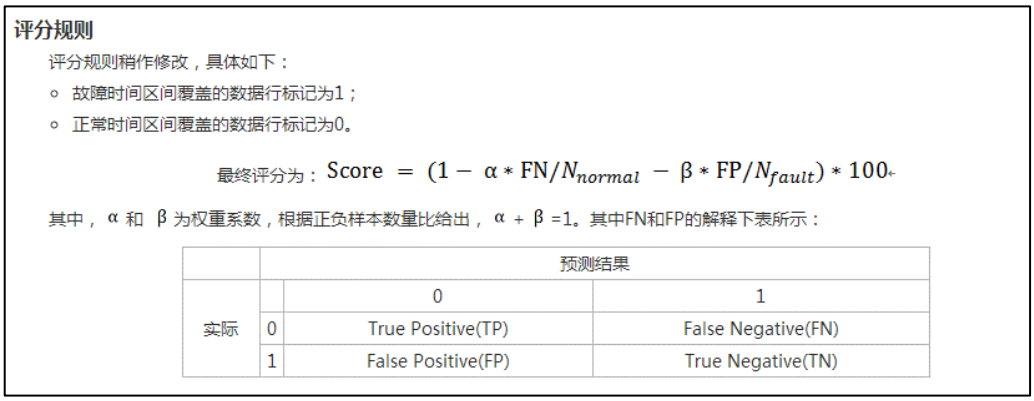
关于风电叶片结冰的文献，大多数论述结冰原因、结冰类型、结冰过程以及结冰给风电场造成的危害。在解决方法上，

一般通过涂料、热风等手段进行防冰除冰，相应开展热传导仿真等工作。通过 SCADA 数据来推测叶片结冰状态，是一个非常创新的思路，如果再结合机器学习算法，已经属于非常前沿的研究了。

参考文献[1]介绍了如何从功率曲线、气象条件、振动等角度判断叶片是否结冰。文中给出的结论是“空气湿度大于 75%、风机功率曲线跌落大于 20%，风机振动的水平加速度大于正常运行的 50%”，预警准确率达到 100%。

这篇论文的思路与此次竞赛十分接近，指明了从哪些数据角度判断结冰的可能性。但文中给出的结论普适性并不高。首先是湿度，它是此次竞赛中并不具备的数据资源；再者，文中所称的准确率其实是机器学习二分类评价指标中的精确率（Precision），是只在预报结冰的样本中的正确比例。可以设想，只要规则设置得越苛刻，触发报警的阈值越罕见，尽量排除那些疑似结冰的状态，只预报十分有把握的情况，那所谓的预警准确率就很容易达到 100%。

相比而言，此次竞赛的评分规则，就显得更加合理。评分公式结合了虚报率和误报率造成的损失，在通过相等或者不等的比例结合起来，得到一个介于 0 到 100 之间的评分，分数越高，效果越好。避免了直接使用准确率的偏颇和漏洞。



评分规则

评分规则稍作修改，具体如下：

- 故障时间区间覆盖的数据行标记为1；
- 正常时间区间覆盖的数据行标记为0。

最终评分为：$\text{Score} = (1 - \alpha * \text{FN}/N_{normal} - \beta * \text{FP}/N_{fault}) * 100$

其中，α 和 β 为权重系数，根据正负样本数量比给出，$\alpha + \beta = 1$。其中FN和FP的解释下表所示：

		预测结果	
		0	1
实际	0	True Positive(TP)	False Negative(FN)
	1	False Positive(FP)	True Negative(TN)

图 3-7：复赛的评分规则

3. 数据解析

通过以上的文献分析，初步找到分析的思路，可以从以下几个角度考虑结冰可能带来的数据影响，并结合建模和数据可视化（以 21 号风机训练数据为例）来探讨特征的适用性。

1、功率

风力发电机组的本质就是通过风产生功率的电气机械设备。而结冰会使得叶片气动外形受到影响，从而降低相同风速下的功率。

下图给出了正常（不结冰）功率散点和通过 bins 区间法拟合的曲线。可以以功率曲线作为正常状态的理论功率，而实际功率距离拟合曲线的距离（残差）作为特征。如果这个残差较大，或者在时间轴上显得很不稳定，可能预示着叶片已经结冰。

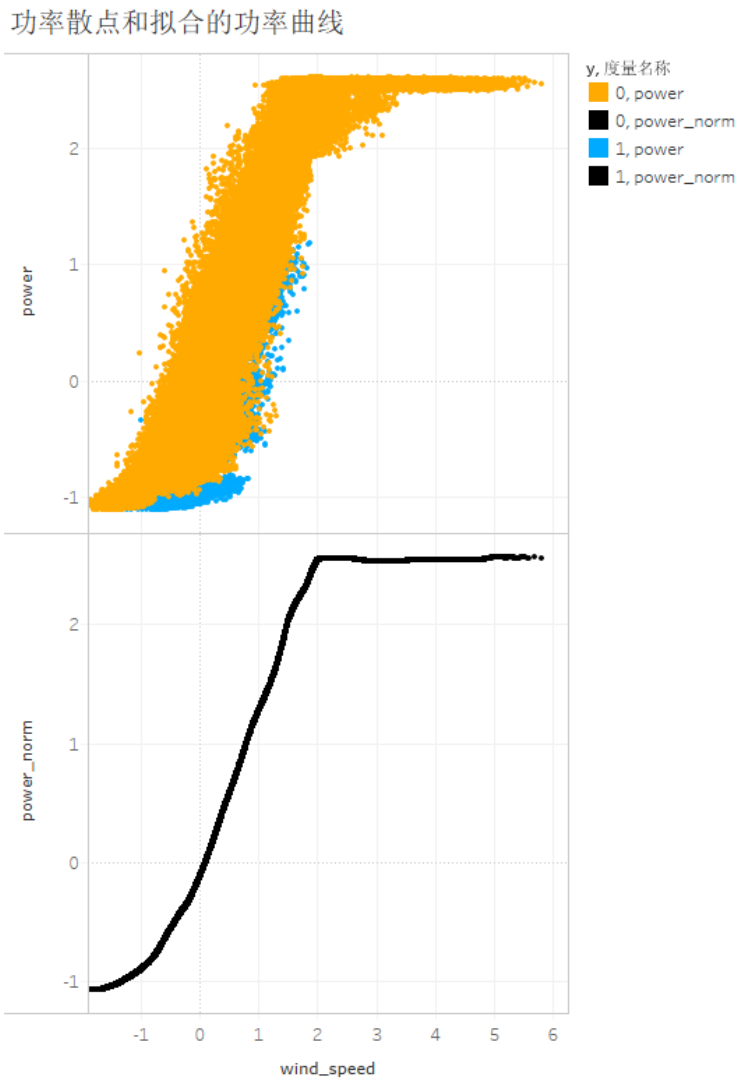


图 3-8：风速-功率散点图

（上图黄色为不结冰数据、蓝色为结冰数据，可见蓝色偏低，且只分布在风速较小区域。

下图为不结冰数据拟合的曲线）

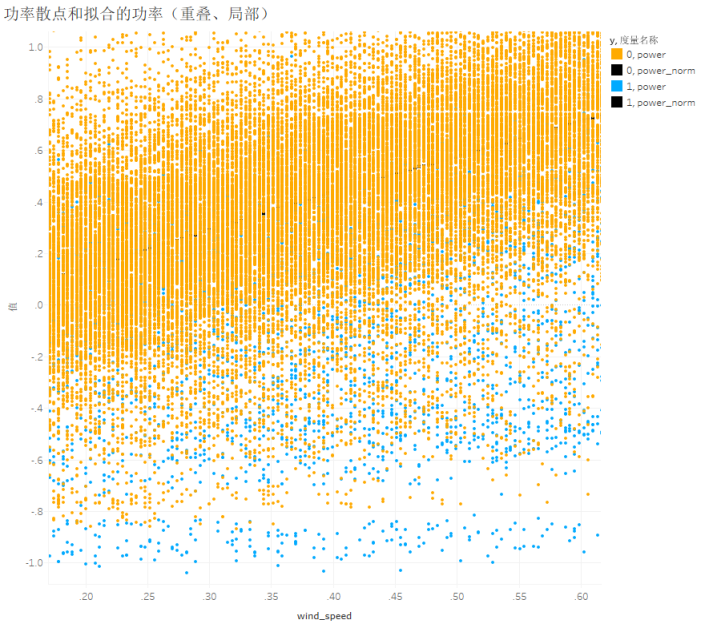


图 3-9：风速-功率散点和拟合的功率

（图 3-8 的局部放大，可见黑色的拟合点，结冰的蓝色散点明显比不结冰的黄色要低，

说明结冰降低了功率，但黄色和蓝色有较多渗透的区域，是造成后期分类精度的难点之一）

2、转速

风机叶片结冰对转速的影响，与功率类似，也会对转速有降低作用。以相同的方法拟合正常转速曲线，计算每个样本的转速残差，并在移动窗口上加以统计。

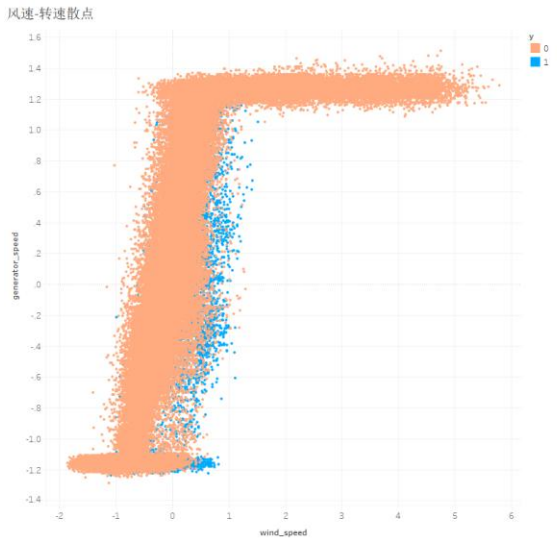


图 3-10：风速-转速散点图

（在相同的风速下，结冰的蓝色点明显比不结冰的红色点要低，说明结冰会降低风机转速）

3、温度

结冰自然与气温有关，进一步的分析发现，将气温和机舱温度做差，更能区分结冰与否的状态。

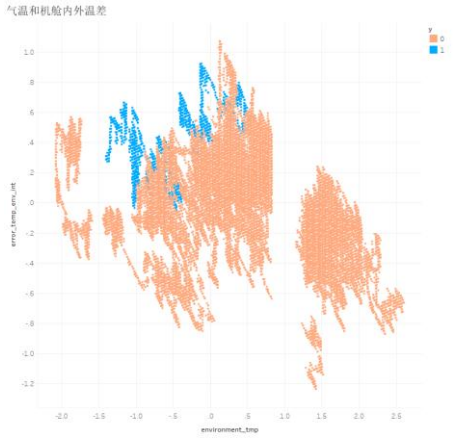


图 3-11：气温和机舱内外温差

（机舱内外温差较好的分离了结冰和不结冰的数据点）

4、风轮不平衡

从机理上讲，正常的风机，三叶片是经过严格的重量和力

矩配平的，使得整个传动链维持较好的平衡状态，避免长期的不平衡造成疲劳磨损。结冰是一个不确定的状态，三叶片结冰程度不可能绝对相同。附着在叶片上的冰块，可能造成风轮在重量和力矩上的不平衡.

不平衡可以从很多参数观察出来，比如带有 pitch 字样的字段来比较 123 叶片的差异。但从后续的建模和数据可视化分析来看，即使在结冰状态，三叶片的桨距角、变桨速率、变桨电机温度等数据都维持较好的一致性，难以从这些字段区分结冰和不结冰。

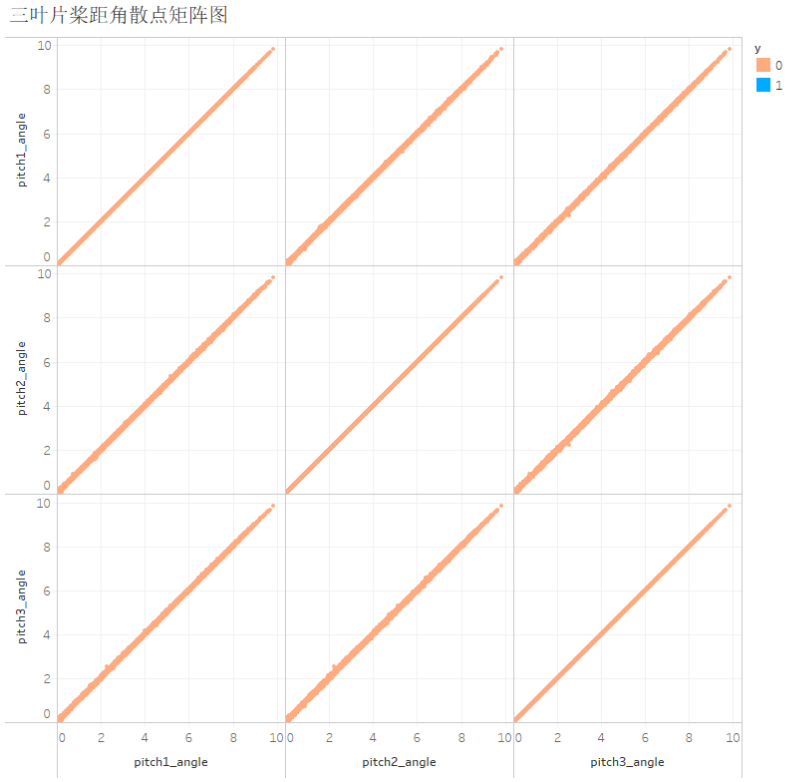


图 3-12：三叶片桨距角散点矩阵图

（三叶片的桨距角基本在同一条 45 度斜线上，说明数值非常接近，相关性非常高，

说明其实三叶片桨距角非常接近，没有区分结冰和不结冰的能力）

5、加速度

一般而言，加速度是是指机舱中两个水平方向（x，y）的加速度传感器得到的数据。由于风机的塔架并不是完全刚性的，风轮若存在不平衡，塔顶机舱的晃动便会加剧。但从后续的建模和数据可视化分析来看，加速度这些字段也难以区分结冰和不结冰。

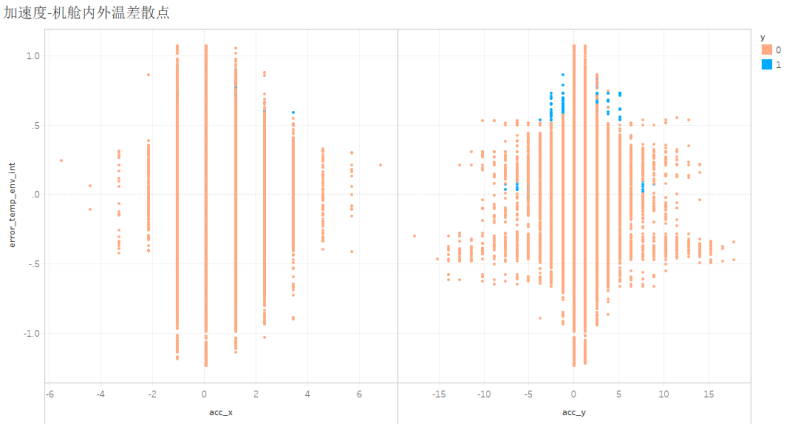


图 3-13：加速度-机舱内外温差散点

（无论是否结冰，加速度散点基本是重叠状态，

预示着加速度可能也无法很好区分结冰和不结冰）

4 方法

方法分析流程:

在数据算法竞赛中，一种普遍的说法是，特征的好坏往往确定了模型的上限，而算法只是逼近这个上限而已。对领域知识的掌握，以及创建良好的特征工程，对于没有足够数学能力独创新算法的工程应用人员来说，显得尤为重要。

首先，通过上述的数据探索和机理研究，构建出大量有利于算法识别的特征。这些特征最初源自领域知识的直觉，后续

经过算法多轮次的数据可视化分析、建模后的特征重要性评估、特征选择以及线上测评后，有一些历久弥新的特征保留下来，它们是:

1、机舱内外的温度差，以及衍生的移动窗口统计特征;

2、风速-功率曲线，以及衍生的移动窗口统计特征;

3、风速-转速曲线，以及衍生的移动窗口统计特征;

4、功率回归模型，以及衍生的移动窗口统计特征;

5、核密度估计，以及衍生的移动窗口统计特征;

6、未来结冰估计。

移动窗口统计特征的含义是，在时间连续的基础上，在每个样本的当下和历史的一定时间范围内，进行算术平均值、中位数、标准偏差、最大值、最小值、上四分位数（75%百分位）、下四分位数（25%百分位）、积分（以 time 为横轴）、趋势（窗口的初值减去终值）等统计指标的计算，并作为特征进入建模当中。

在整个特征工程优化过程中，我们团队特别重视“特征监控”这个概念。由于训练集只有两台风机的数据，每台风机由于安装位置和部件制造差异性，存在或多或少的差异。相互作为验证集之后，有些特征在本地是有效果的，但到了第三台风机上就不一定有效果。所以要特别重视特征的普适性。

然后，在算法上采用先简单、后复杂的原则。首先尝试一些基础算法，例如逻辑回归、Lasoo 回归、岭回归，设定 Baseline。再利用竞赛圈内广受好评的决策树集成算法，例如

随机森林、XGBoost、LightGBM 做为最主要的算法，参考竞赛说明，设定自定义的评估函数。引入自动化调参工具 Hyperopt，利用验证集效果调节超参数，使得算法表现最优。重要模型都是特征选择之后的特征子集，来构建最终模型。

最后在算法融合上，尝试了 Bagging、Blending、Stacking 等经典方法。但由于训练集风机数量太少，在本地有所提升的融合结果，并不一定能在线上获得提升。实际上，功率回归模型、未来结冰估计等方法，是通过建模构建新的特征，然后再叠加到下一次的模型中，本质上非常类似 Stacking 的思想。

总之，特征工程为主，算法优化为辅，这是我们团队贯穿于整个竞赛的思路。

方法的独到之处:

我们团队方案的特色主要在于特征构建方面，参考了大量的领域知识，灵感来源都是数据背后的运行机理。此外，我们团队提出两种非常有特色的特征构建方法：未来结冰估计和核密度估计特征。

未来结冰估计特征：赛题本身是要求参赛者根据当下和历史的数据，推测当下的结冰状态。在实际研究过程中，我们发现很难抓取到结冰的变化趋势。如果通过未来严重结冰时期，回溯、推测当下的结冰状态又是违规的，在实际部署中也是不可实现的。

直接用未来数据是不可以的，那使用训练数据模型建一个

模型预测未来总是可以的。具体来说，建立这么一个模型，根据当下和历史的数据，推测未来一段时间的结冰状态。模型建立好之后，构建未来结冰估计概率的特征，对未来做出推断，增强主模型（预测当下是否结冰）对变化趋势的敏感度，有利于提高预测精度。

核密度估计特征：核密度估计可以简单理解为对直方图的光滑拟合。如果结冰和不结冰的分布差异较大，可能呈现双峰分布，那结冰的样本对于某个正常情况下的核密度估计，就是处于较低的密度数值。也就是说在这个特征上，分布上处于比较罕见的情况。既然是罕见情况，就有可能是出现了异常。算法可以结合其他特征的情况，从而判断可能是发生了结冰。

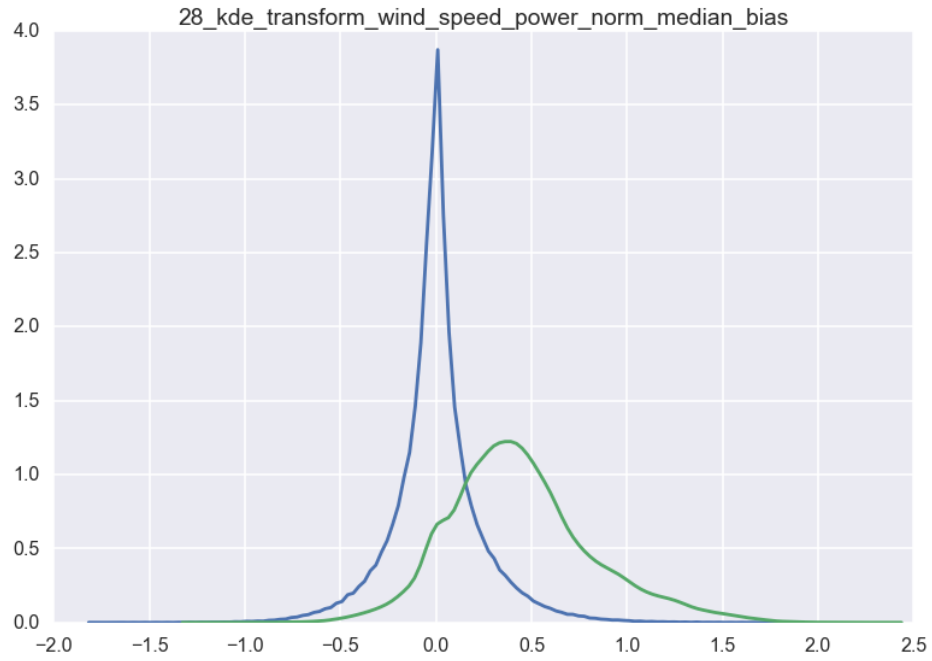


图 3-14：风速-功率拟合曲线与实际功率的残差的核密度分布图（蓝色为不结冰数据，基本分布在 0 的附近，而且非常集中。

说明不结冰的数据，趋于正常，有较高的集中程度。

而绿色曲线是结冰数据，偏离 0 位较多，且较为分散，说明结冰的程度有所不同，并不能像正常状态那样集中）

5 验证

对该数据的验证、与其他方法的对比

提交编号	提交日期	结果名称	10、14号风机最终测试集得分	08号风机测试数据集得分	验证集平均得分	15号训练21号验证验证集得分	21号训练15号验证验证集得分	特征数量	备注
Z5	20170824	submit_test_06_LGB_fb03_rol80_feture_fs_nozero_top1_bagging_2cv_wt_test081014	-	85.5807	84.4420	83.1985	85.6855	5	未加入未来结冰估计特征
F1	20170908	final_1st_16_LGB_fb08_rol80_future_nrol20_fs_11+nrol20_top5_bagging_2cv_wt_type_ab0.5_test08_10_14	76.1252	87.4201	85.0216	82.8284	87.2147	27	包含未来结冰估计特征
F2	20170922	final_2nd_27_LGB_fb11_kde_rol80_future_nrol5_10_20_40_80_fs_kw_nozero_top1_bagging_2cv_wt_type_ab0.5_test08_10_14	76.3538	88.3508	84.4308	83.0346	85.8270	103	加入核密度估计特征

上表列举了三次提交记录的得分情况。可以看到加入未来结冰估计特征和核密度估计特征之后，08 号风机得分有 2 分和 1 分的提升。在 10、14 号风机上也有相应提升。

6 结果分析与经验总结

经过决赛答辩后，我们团队了解了其他优秀团队的竞赛方案，逐渐认识到我们团队在此次比赛中存在的一些不足之处:

首先是在数据探索、可视化分析上做得不够精细。由于对领域知识抱有充分的自信，难免产生一定依赖性。基本上所有的方法是以一种头脑风暴的形式，构建大量可能有效的特征，然后让算法去识别和筛选有用特征。而不太注重从数据本身出发，探索一些相关规律。

其次，上述这种简单粗暴的方法，也会加重对硬件资源的依赖性。相比于其他团队只使用普通的笔记本电脑，我们团队则使用配置几十线程的至强 CPU，上百 G 的内存的台式工作站进行特征构建和迭代调参，所用的方法比较依赖于宽裕的计算资源。

最后，在算法本身上，我们团队并没有做太多的探索和调整。比如设定更加优良的目标函数（损失函数），写出 XGBoost 和 LightGBM 中梯度下降的一阶导数和二阶偏导数构成的海森矩阵。

计算速度，使用的硬件:

工作站 1:

处理器英特尔 Xeon E5-2667 v2 @ 3.30GHz 8 核 16 线程 (X2)

内存 192 GB (DDR3 1333MHz)

工作站 2:

处理器英特尔 Xeon E5-2683 v3 @ 2.00GHz14 核 28 线程 (X2)

内存 128 GB (DDR4 2133MHz)

在计算速度上，特征构建和迭代调参占据了大部分运行时间。从竞赛官网下载的原始文件到最终的提交文件，实际运行时间可能长达数天。

7 参考文献

[1]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

（四）基于数据驱动和非均衡数据学习的故障预测研究

1.团队介绍

团队名称：sugerlin

成员姓名	团队角色	职位
林文芳	队长	北邮研究生

林文芳，北京邮电大学信息与通信工程学院学生，在泛网无线教育部重点实验室之移动生活与新媒体实验室（MINELAB）做学术研究，研究方向包括物联网与工业互联网、工业大数据、机器学习、时间序列分析等；吴振宇，移动生活与新媒体实验室研究生导师，主要从事物联网技术与服务、工业大数据技术与服务、情景感知与智能服务系统等科研和教学工作。

2.背景简介与文献调研

随着人工智能技术和工业的发展，越来越多的传感器数据需要在工业过程中收集与分析，故障预测也开始受到学术和工业领域的关注。故障预测的成功在于不但使工业系统做到节能降耗，同时还能预防一些重大事故的发生。一个稳定准确的故障预测系统能够防止重大事故、节约成本以及提高制造效率。然而，现代工业系统的复杂性严重阻碍了我们对系统结构的认知。因此，数据驱动型的方法更适宜应用到工业领域的故障预测问题。

同时，在全球信息化的不断发展下，不管是在工业化生产、数字化管理还是我们的日常生活中，随着信息化发展所带来的知识爆炸，非平衡数据集分类问题也越来越受到研究者的关注。

例如，在工业中，很多设备长时间都处于一个正常的状态，有极少部分的时间处在故障状态，这使得我们收集到非平衡数据。而且，少数类信息才是我们关注的重点，因此，对非平衡数据分类问题的研究也有着广泛的应用前景和实际意义。

关于工业故障预测方面有很多值得借鉴的论文，例如 Isermann 等人[1]使用传统的基于模型的方法通过相关经验和专业知识来识别和预测故障，Yin 等人[2]设计了一个数据驱动的故障识别系统来识别风机的故障。也有许多机器学习算法用于故障分类，例如 KNN[3]、SVM[4]、RF、GDBT[5]和 XGBoost[6]等算法。同时针对于非均衡数据集，除了随机过采样和降采样的方法外，SMOTE[7]也是一种有效的方式来合成少数类样本。同时，Xu 等人[8]提出 EasyEnsemble 和 BalanceCascade 两种集成算法来解决非均衡数据集的问题。最后，我们根据 PHM 大赛中的两篇论文[9][10]提出的特征提取方式和故障预测方法来解决我们遇到的问题。

3.数据解析

首先我们对数据进行的简单的初步统计分析。我们观察故障持续的时间以及整个数据集中正负两类的特点。图一描述了所有结冰时间段持续时间的分布，横坐标是结冰持续的时间，纵坐标是结冰时间段的个数。从图 3-15 中我们可以看出，结冰故障持续时间 90%左右都发生在 140 分钟以内，最短的是 12 分钟，最长的是 1933 分钟，这个对我们接下来的特征选取以及后期的时间窗选择有很大的影响。

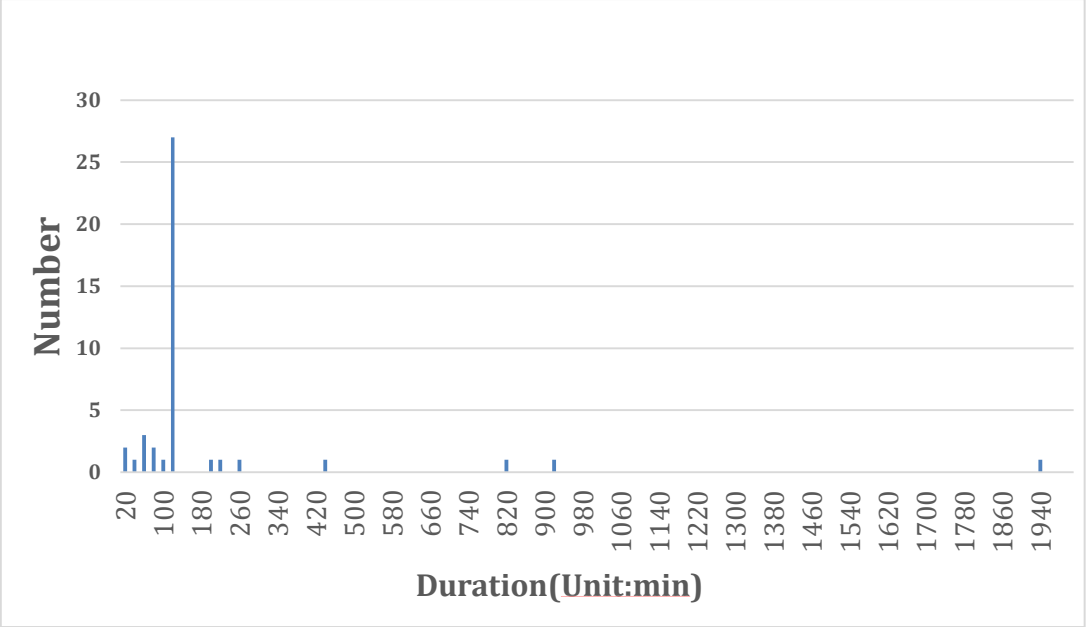


图 3-15：结冰故障持续时间统计

图 3-16 描述的是给定的训练集中正常时刻个数与故障时刻个数所占整个数据的比例，15（0）表示训练集 15 号风机处于正常时刻所占的百分比，15（1）表示训练集中 15 号风机处于结冰故障时刻所占的百分比，21（0）表示训练集中 21 号风机处于正常时刻所占的百分比，21（1）表示训练集中 21 号风机处于结冰故障时刻所占的百分比。从图中我们可以看出，训练集中正常和结冰故障的时刻个数比例接近 1：16，是个较为不均衡的数据集，这要求我们后期选择合适的分类器以及对此原始数据进行数据层面的处理有一定的影响。

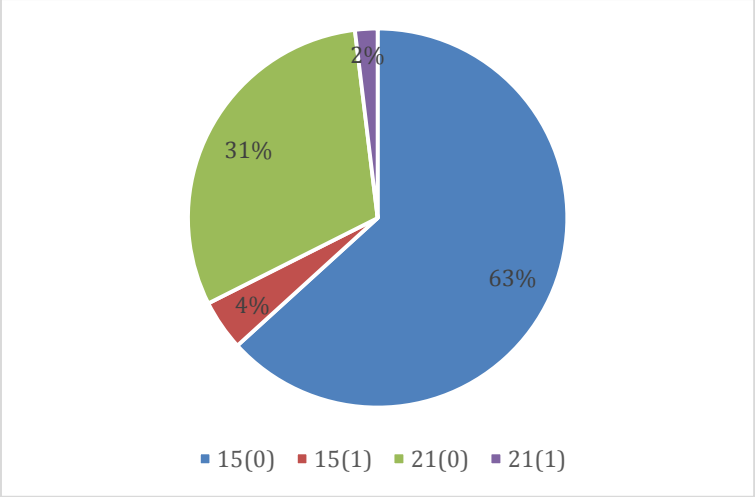


图 3-16：结冰和正常时刻个数比例

4. 方法

（1）问题分析

首先，我们简单对问题进行阐述，如图 3-17 所示。我们需要根据风机在连续时间内的 SCADA 原始数据，也就是每一时刻均对应 27 个连续数据变量（风机的工况参数、环境参数以及状态参数）来预测风机在何时发生结冰故障，从而预测结冰故障发生的起始时间和终止时间。我们在不了解风机叶片的结构或物理性质的情况下，采用数据驱动的方法。更具体地说，我们从原始数据中提取若干无序特性，将时间序列切片化。最后，我们对分类器进行训练，从而预测故障发生的时间。

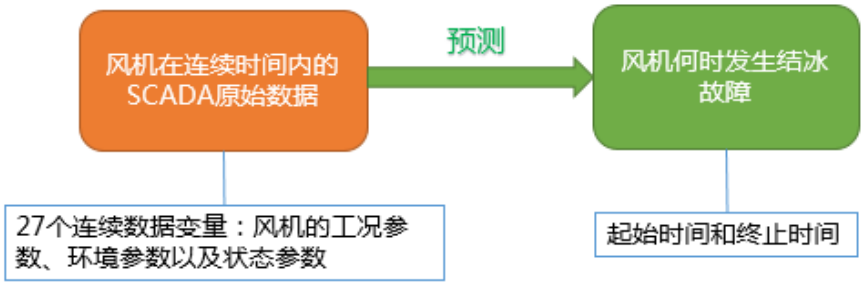


图 3-17：风机叶片结冰预测问题分析

面对这一挑战，我们需要对测试数据集中的原始数据是否存在结冰故障进行预测，从而预测发生结冰故障的时间段，包括起始时间和终止时间。我们通过一下三个步骤将此预测问题转化为分类问题：（1）切割；（2）分类（3）结果处理。具体步骤如图 3-18 所示。第一步，将基于时间序列的测试数据$\boldsymbol{S_{raw}}$按照时间窗为 K，步长为 L 切割成若干时间片段$\boldsymbol{p_i = [t_i, S_i]}$，每个时间片段均包括一个时间戳和其时间窗内包括的所有传感器数据信息。第二步，对此若干片段进行分类，预测若干时间片段对应的标签（故障为 1，正常为 0）。第三步，如图 3-19 所示，将有相同标签的相邻时间片段合并，同时消除故障持续时间较短的片段，最后得到若干故障时间段，即预测故障发生的时间。

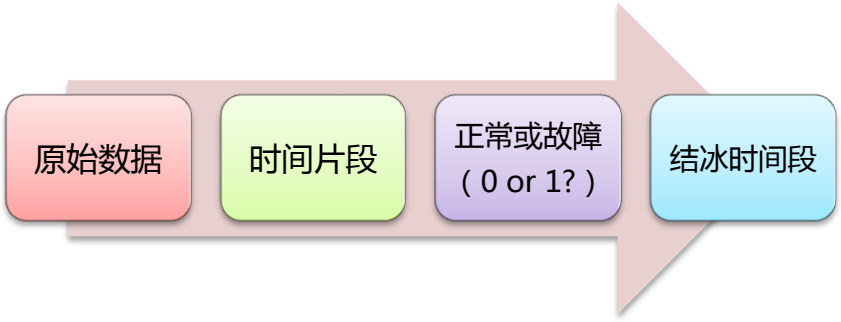


图 3-18：问题的理解——时间片段的分类问题

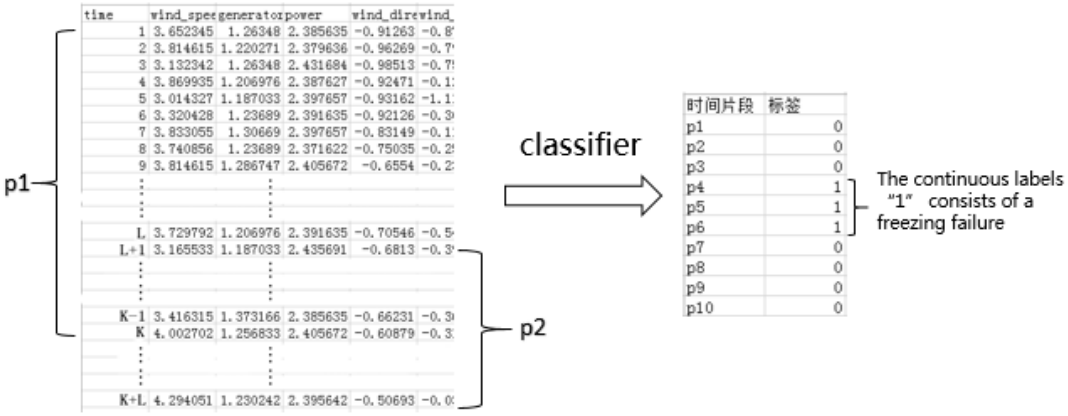


图 3-19：结果处理

（2）数据预处理

数据预处理主要包括两部分:

采样时间标准化。原始数据的采样时间不统一，选择采样间隔 7 秒，删除采样间隔不为 7 秒的数据。

提取有效数据。根据题目要求，删除不包括在*_normalinfo.csv 和*_failureinfo.csv 两个文件的时刻。

3.特征加工与特征选择

这一节中，我们介绍特征提取的方式。如图六所示，原始数据中共有 26 个传感器或执行器的数据信息，将基于时间序列的测试数据$\boldsymbol{S_{raw}=[S_1,S_2,\dots,S_n]}$（$\boldsymbol{S_1=[t,x_1,\dots,x_{26}]}$）按照时间窗为 K，步长为 L 切割成若干时间片段$\boldsymbol{p_i=[t_i,S_i]}$，每个时间片段均包括一个时间戳和其时间窗内包括的所有传感器数据信息，切割方式如图 3-**20** 所示。我们选取 K=106，L=20，得到若干时间片段。所以每一时刻的特征$\boldsymbol{f_t=[x_1,x_2,\dots,x_{26}]}$，那么每一个时间片段的特征$\boldsymbol{p_{i,t}=[f_t,f_{t+1},f_{t+2},\dots,f_{t+105}]}$，因此形

成一个1 × 2756的矩阵作为特征向量。

	A	B	C	D	E	F	G	H
1	time	wind_spee	generator	power	wind_dire	wind_dire	yaw_posit	yaw_sp
2	2015/11/1 20:20	1.859993	1.223595	2.51579	-2.07274	-2.07363	-0.65534	0.030
3	2015/11/1 20:20	1.911625	1.293394	2.313551	-2.01059	-1.61514	-0.65534	0.030
4	2015/11/1 20:20	1.635027	1.280099	2.507799	-2.05375	-0.28274	-0.64957	0.170
5	2015/11/1 20:20	1.786234	1.280099	2.349593	-2.00714	-2.23448	-0.65534	-0.00
6	2015/11/1 20:20	1.786234	1.26348	2.321566	-2.26437	-1.42896	-0.63792	0.415
7	2015/11/1 20:20	2.022264	1.286747	2.389643	-2.17805	-0.79316	-0.62627	0.623
8	2015/11/1 20:21	2.202974	1.280099	2.455728	-1.82587	-0.02057	-0.59132	1.077
9	2015/11/1 20:21	2.298861	1.416375	2.483755	-1.29588	0.86854	-0.59132	0.853
10	2015/11/1 20:21	2.60865	1.256833	2.503791	-1.0093	0.344194	-0.59132	0.623
11	2015/11/1 20:21	2.523827	1.256833	2.507799	-0.49658	1.055988	-0.59132	0.484
12	2015/11/1 20:21	2.39106	1.243537	2.513798	0.009247	1.833642	-0.59132	0.344
13	2015/11/1 20:21	2.103399	1.23689	2.51579	0.311359	1.689257	-0.59132	0.247
14	2015/11/1 20:21	2.254606	1.243537	2.511806	0.463279	0.848276	-0.59132	0.170
15	2015/11/1 20:21	2.000136	1.157118	2.48577	0.42012	-0.30934	-0.59132	0.138
16	2015/11/1 20:21	1.81205	1.343252	2.339587	0.504711	0.317596	-0.59132	0.107
17	2015/11/1 20:22	2.328365	1.323309	2.487762	0.682526	1.986893	-0.59132	0.068
18	2015/11/1 20:22	2.379997	1.280099	2.503791	0.88451	0.896404	-0.59132	0.030
19	2015/11/1 20:22	2.582835	1.200328	2.503791	0.812003	-0.0459	-0.5855	0.205
20	2015/11/1 20:22	2.184534	1.26348	2.517805	0.938027	1.627196	-0.59132	-0.00
21	2015/11/1 20:22	2.354181	1.223595	2.505783	1.129652	1.528406	-0.59132	-0.00
22	2015/11/1 20:22	2.324677	1.300042	2.513798	1.164179	1.011659	-0.59132	-0.00
23	2015/11/1 20:22	2.549643	1.293394	2.507799	0.998449	0.284666	-0.5855	0.170
24	2015/11/1 20:22	2.586523	1.280099	2.507799	0.931121	0.139015	-0.59132	-0.00
25	2015/11/1 20:23	2.306237	1.187033	2.507799	0.910405	0.046557	-0.59132	-0.00
26	2015/11/1 20:23	2.402124	1.243537	2.503791	0.894868	0.106085	-0.59132	-0.00
27	2015/11/1 20:23	2.379997	1.300042	2.521812	0.768844	1.172509	-0.5855	0.170

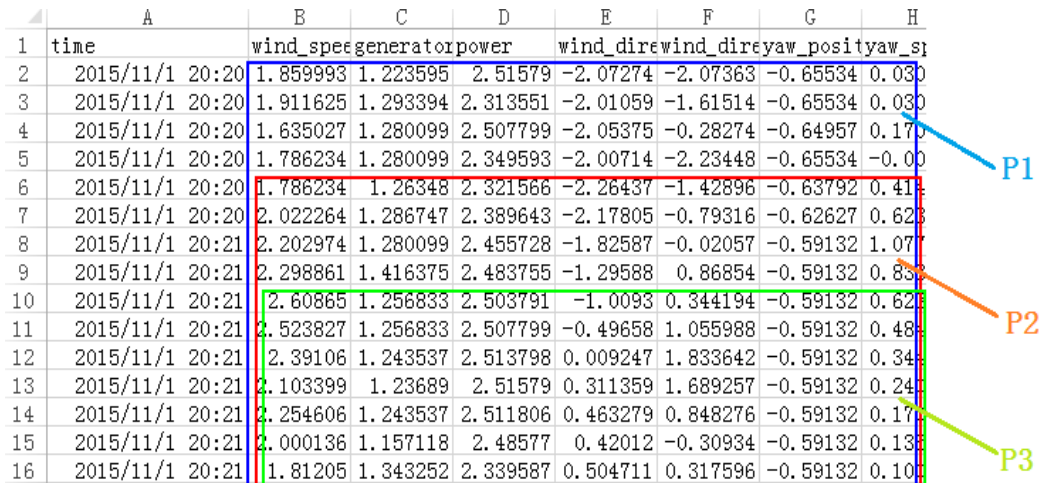


图 3-20：特征加工与特征选择的方法

5. 验证

首先，我们先来设置实验。如图 3-21 所示，我们设置实验的训练集、验证集和测试集。验证集为 21 号风机第 90000-149999 时刻，训练集为全部的 15 号风机时刻和 21 中除于验证集部分的其他数据，测试集为 08、10、14 号风机。

图 3-21：实验设置

根据上文的数据分析，针对于非均衡数据集，我们选择对

此非均衡问题不敏感的算法来进行学习。同时，为了保证稳定性，我们选择集成模型平均的方法，最后使用精确率、召回率和 F 值来评价方法的性能。我们首先用 RF、GBDT、XGBoost 三种基分类器中进行比较，如图 3-22 所示，我们在三种不同的基分类器下，对分类结果的准确率、召回率、精确率和 F1 值进行对比，我们发现 XGBoost 的召回率最高，可以更好的识别出结冰故障时间片段，精确率较高，在保证有较多结冰故障识别出来的前提下，又能保证有较高的精确率。F1 值是精确率和召回率的综合评价指标。因此，我们最后选择 XGBOOST 作为基分类器。

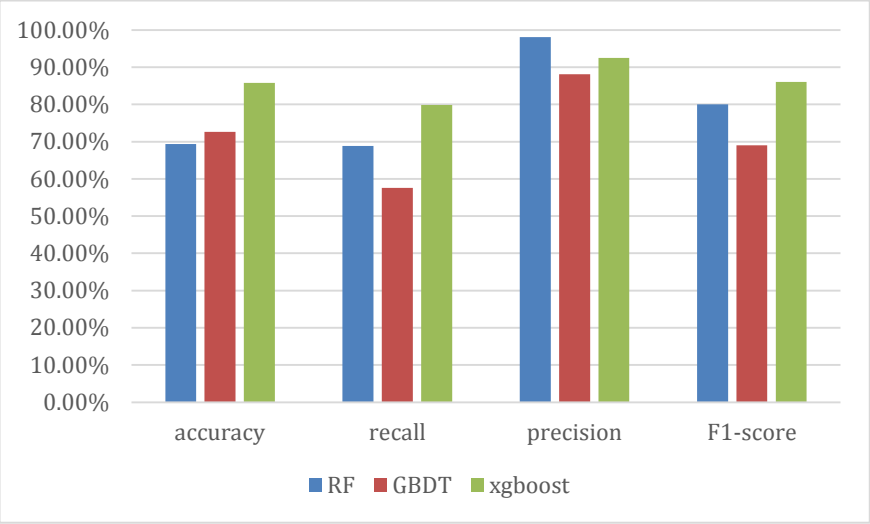


图 3-22：RF、GBDT、XGBoost 三种基分类器在不同评价指标下的比较

在选定基分类器后，我们在训练集上做五折交叉验证来检测模型的性能。将训练集分成 5 个子集（每个子集中国足航和正常的比例相同），每个样本单独作为测试集，其余 4 个作为初步训练集。由于数据的非均衡特点，我们使用随机过采样、降

采样和调整分类器阈值两方面对结果进行处理。我们设置初步训练集中故障与正常的比例，按照比赛给定的打分机制在验证集上计算得分，最后得分对应的故障与正常比例即为最优比例。如图 3-23 所示，我们发现在比例为 1：1.4，阈值为 0.5 的时候在验证集上有最高的得分，因此在接下来的实验中，我们均设置故障与正常的样本个数比例为 1：1.4。

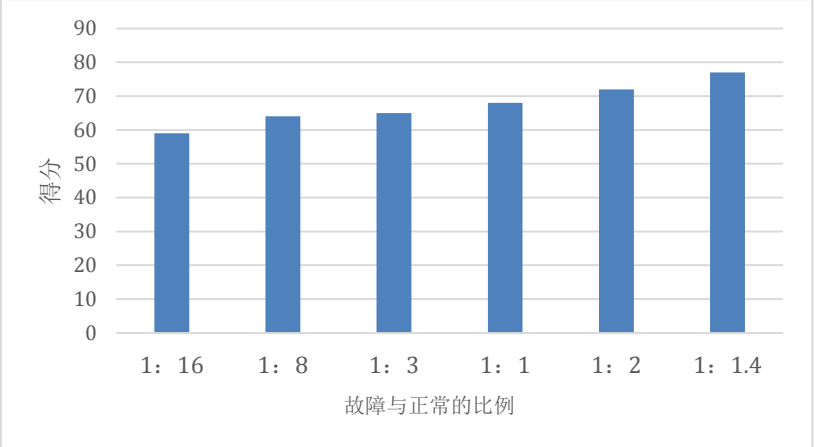


图 3-23：不同故障与正常样本个数比例下的验证集得分

在故障与正常最优比例下，为了保证模型的泛化性能，我们做交叉验证来选择性能最好的模型应用于最终的测试集上。我们计算交叉验证模型在初步测试集的分类效果和验证集上的得分，如图 3-24 所示。由于得分受故障识别率和正常识别率的双重影响，我们发现，两个识别率都高的情况下，在验证集上有更高的得分。因此，我们选择序号 5 作为训练集建立模型，并通过 XGBoost 参数调优和结果处理来对得分进行一个小幅度的提高。

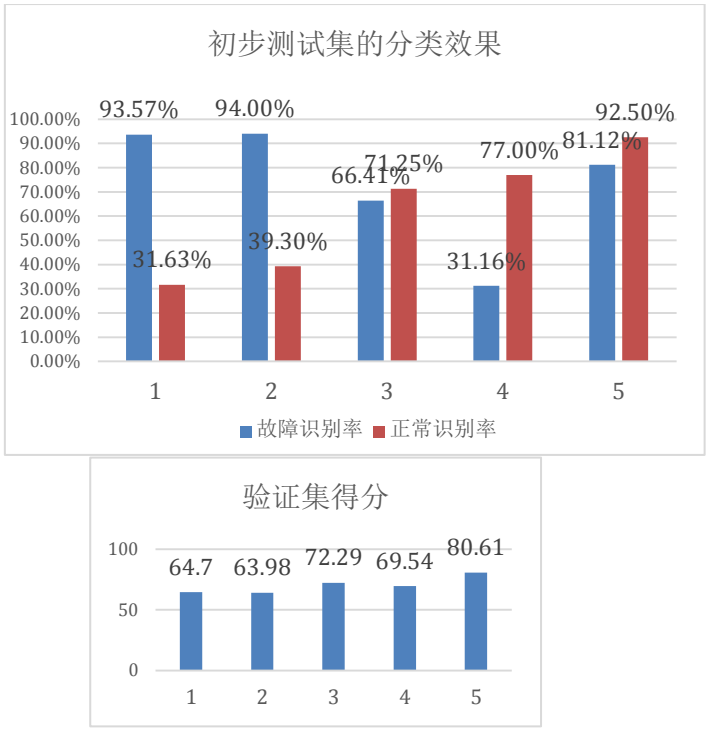


图 3-24：交叉验证在初步测试集和验证集上的结果

XGBoost 参数调优：

maxdepth	eta	threshold	score
6	0.3	0.5	74.808
6	0.3	0.5	75.102
2	0.03	0.5	76.083
2	0.03	0.4	76.236

结果处理（删除分类噪音片段、合并相邻故障段间隔时间短的故障片段）：

before	76.236
after	77.087

最后在 08 测试集上进行测试：

08 测试集	77.087
10，14 测试集	76.235

6. 结果分析与经验总结

1. 经验教训

（1）做实验前，数据分析是至关重要的，我们需要通过对数据的初步分析来认知数据特点，从而正确分析问题解决问题。

（2）基分类器的选择至关重要，需要做一些基础实验来选择合适的基分类器来做深入实验。

（3）一般的分类器无法对非均衡数据进行很好的分类效果，因此需要对非均衡问题进行分析，无论是从数据层面还是算法层面。同时，对于非均衡数据集，评价方法也与通常使用的准确率不同。

（4）实验很重要，需要多做实验，横向纵向比较来发现问题并解决问题。

2. 硬件环境

我们在 Ubuntu14.04 LTS 的阿里云服务器上进行数据处理、特征选择以及模型验证和测试过程。主要使用的编程语言为 JAVA 和 Python。

3. 可改进的地方

（1）我们使用数据驱动型方法来解决问题。接下来可以适当加入物理模型，了解风机结冰的原因来分析问题。

（2）特征提取上我们只是对原始数据进行了时间窗的切割处理，后期可能需要对特征选择上进行改进。

（3）对于非均衡数据集，后期应该使用数据层面或者算法层面的方法来解决非均衡数据的问题。

7. 参考文献

[1] Isermann, R. (2005). Model-based fault-detection and diagnosisCstatus and applications. Annual Reviews in control, 29(1), 71-85.

[2] Yin, S., Wang, G., and Karimi, H. R. (2014). Data-driven design of robust fault detection system for wind turbines. Mechatronics, 24(4), 298-306.

[3] He, Q. P., and Wang, J. (2007). Fault detection using the knearest neighbor rule for semiconductor manufacturing processes. Semiconductor manufacturing, IEEE transactions on, 20(4), 345-354.

[4] Samanta, B. (2004). Gear fault detection using artificial neural networks and support vector machines with genetic algorithms. Mechanical Systems and Signal Processing, 18(3), 625-644.

[5] 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.

[6] Tianqi Chen and Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2016, pp. 785-794

[7] N. V. Chawla, K.W. Bowyer, L. O. Hall, andW. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002

[8] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, Exploratory Undersampling for Class-Imbalance Learning, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. 39, NO. 2, APRIL 2009.

[9] Hyunjae Kim, Jong Moon Ha, Jungho Park, Sunuwe Kim, Keunsu Kim, Beom Chan Jang, Hyunseok Oh and Byeng D. Youn，Fault Log Recovery Using an Incomplete-data-trained FDA Classifier for Failure Diagnosis of Engineered Systems，2015 PHM Data Challenge Competition.

[10] Cong Xie, Donglin Yang, Yixiang Huang, and Donglai Sun，Feature Extraction and Ensemble Decision Tree Classifier in Plant Failure Detection，2015 PHM Data Challenge Competition.

（五）基于敏感特征的风机叶片结冰预测算法

1.团队介绍

团队名称：XJTU_DL

成员姓名	团队角色	职位
雷亚国	技术顾问	教授
李宁波	组长	博士研究生
闫涛	组员	博士研究生
郭亮	组员	博士后

XJTU_DL 团队的三位成员和指导老师均来自西安交通大学机械工程学院。团队指导教师雷亚国教授长期从事机电设备智能诊断和寿命预测研究，是首批国家优秀青年科学基金获得者、教育部“长江学者奖励计划”青年学者、中组部“万人计划”青年拔尖人才，为本竞赛的顺利开展提供了有力的技术指导。团队成员包括博士后郭亮、博士生李宁波和闫涛，队长为博士生李宁波。团队成员简介和具体分工如下：

李宁波：2015 年本科毕业于四川大学制造科学与工程学院机械设计制造及其自动化专业，同年保送至西安交通大学机械工程学院攻读长学制博士学位，目前研究方向为基于深度学习的机械设备剩余寿命预测。在本次比赛中担任队长，负责算法设计与开发；

闫涛：2016 年本科毕业于中南大学机电工程学院机械设计制造及其自动化专业，同年保送至西安交通大学机械工程学院攻读长学制博士学位，目前研究方向为基于衰退模型的机械设备剩余寿命预测。在本次比赛中，负责工业背景调研；

郭亮：2011 年本科毕业于西南交通大学机械工程学院测控技术与仪器专业；2016 年博士毕业于西南交通大学机械工程学院机械电子工程专业，同年进入西安交通大学机械工程学院从事博士后研究，目前研究方向为机械设备状态监测、智能故障诊断和剩余寿命预测。在本次比赛中，负责方案设计与数据预处理。

2.背景简介与文献调研

风电是目前最成熟、最具发展潜力且基本实现商业化的新兴可再生能源技术。在全球各国中，中国的风电发展举世瞩目，年新增风电装机占全球的比例从 2006 年的不足 10%上升到 2010 年的 49%，且保持着不断增高的趋势。但风能获取的特殊性决定了大量风机需布置在高纬度、高海拔的寒冷地区。以我国为例，至 2016 年，全国风机累计装机总量的 65%以上位于北方寒冷地区[2]，如图 1 所示。而工作在寒冷地区的风机受霜冰、雨凇和湿雪等气象条件影响，极易发生叶片结冰现象，进而引发一系列后果。具体来说有以下危害[3]：1）结冰后的风机叶片翼型发生改变，导致风能捕获能力下降，加之叶片上附着冰层，增大了叶片转动所需能量，最终导致风机发电功率损耗；2）风机叶片结冰后，直接导致叶片部分结构参数改变，继而影响其固有模态参数，诱发叶片断裂；3）当风机叶片结冰积累到一定程度后，冰层受自重影响断裂飞出，极易击中风场巡检人员，造成人身事故。

由上可见，倘若不能及时检测并消除叶片结冰故障，轻则会导致风电设备损耗加剧、影响风场经济效益；重则将导致风机叶片断裂，甚至带来机毁人亡的惨剧。鉴于此，无论是为了降低风机的维护成本、提高风场经济效益，还是为了降低运行风险、提高风机的安全系数，都需要大力发展风机叶片结冰检测技术，使其充当风机运维的“安全卫士”，降低维护维修成本、避免重大事故发生。当前，国内外众多研究机构和企业对风机叶片结冰现象的检测技术进行了深入研究。根据使用方法的不同，风机叶片结冰检测方法主要分为三类：机理模型方法、数据驱动方法和数模联动方法。

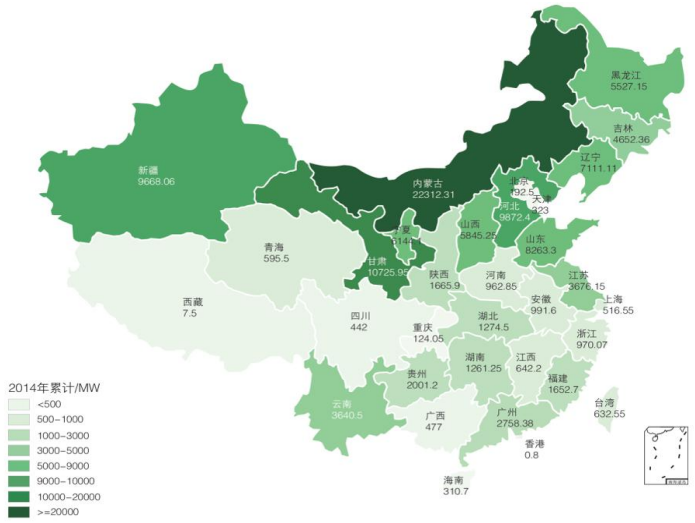


图 3-25：2016 年中国各省（区、市）累计风电装机容量[19]

基于机理模型的方法主要通过对叶片结冰机理、结冰风机动态响应和风机输出功率等内容进行理论分析研究，建立叶片结冰判断模型并根据监测数据判断当前时刻风机叶片是否发生结冰。而这些方法可划分为直接法和间接法[4]。其中直接法使

用光纤监测、红外检测、超声检测[5-7]等手段，首先根据风机运行环境设计模拟实验，采集叶片质量、叶片光反射性、热传导性、介电常数等指标在正常和结冰状态下的数值，而后根据专业人员的经验设定阈值判断叶片是否结冰[4]。而间接法则以环境温度和湿度、风机功率损耗[8-9]等为研究对象，根据其与叶片结冰程度间的相互关联，建立相应的结冰检测模型[8]。由于基于模型的检测方法能够借助于研究人员的专业知识，在分析建模准确的情况下，能够高效、准确地判断叶片健康状态，因此现有商用风机状态监测系统中大多使用基于模型的叶片结冰检测方法[10]。但基于模型的检测方法也存在着不可忽视的弊端：1）严重依赖研究人员的专业知识；2）模型通用性不足，当工作环境等因素发生变化时，检测准确率将大打折扣；3）绝大多数直接检测方法需额外布置传感器，成本偏高[4]。

基于数据驱动的方法利用智能模型直接从监测数据中通过智能模型挖掘可以表征风机叶片结冰的有效信息，从而判断风机叶片的结冰状态。随着 SCADA 系统在风电行业的广泛使用[11]，研究人员可以获取风机各部件或子系统的大量状态监测数据，而异常运行状态信息往往蕴含于其中。倘若使用合适的分析方法和数据挖掘手段，则可充分利用已有的大量监测数据，判断风机是否发生故障[12]。这一类方法能从已有监测数据中挖掘故障信息、无需过分借助专业知识，可被概括为数据驱动方法。迄今为止，数据驱动方法在风机齿轮箱、发电机和主轴等零部

件的故障检测中[13,14]，并已有相应的状态监测系统问世[15-17]。而当前尚未见到使用数据驱动方法对叶片结冰进行检测的文献和商用系统，但也可借鉴已有文献中的思路，以风机 SCADA 数据为基础，挖掘运行状态信息并据此判断叶片是否结冰。

基于数模联动的方法综合利用机理模型的可解释性和数据驱动模型强大的数据表征能力，在机理分析的基础上通过智能模型判断风机叶片的结冰状态。目前，数模联动方法还处于早期研究阶段，相关研究成果较少。文献[18]通过分析风机的运行状态变量与对应产生的叶片扭矩值关系构建风机叶片结冰的特征值，并通过故障树方法判断风机叶片结冰状态。文献[19]提取了风机机舱振动和风机的功率曲线作为风机叶片结冰的特征值输入决策树进行叶片结冰的监测并进行早期除冰决策。虽然数模联动方法在风机叶片结冰检测和预测具有一定的优势，但是因其发展时间较短，目前工业界还未大规模应用。但随着人们对风机叶片原理理解的加深和智能科学的进一步发展，将逐步推动这一研究和产业方向的发展。

本小组此次竞赛方案采用数模联动方法。在充分分析风力发电机原理及叶片结冰对监测变量影响的基础上，提取了风速、网侧有功功率、环境温度、机舱温度、风速和功率非主成分方向投影 5 个显性特征构成特征向量。提取的特征向量输入逻辑回归分类器进行风机叶片结冰预测。

3.数据解析

目前风机运行的实时数据主要由 SCADA 系统进行存储，SCADA 系统中存储的数据通常有上百个变量。本次竞赛的数据来自某风场内 5 台风机，数据采集历时 2 个月，总计采集了一百多万个时间戳的 SCADA 监测数据。其中每个时间戳的 SCADA 数据包括 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，数据概况如表 3-4 所示。

表 3-4：比赛数据概况

类型	风机编号	数据时间范围	时间戳数	组数
训练	15	2015.11-2015.12	393836	3854
	21	2015.11	190494	1855
测试	08	2015.11	202328	1982
提交	10	2015.11	174301	
	14	2015.11	163732	

由表 3-4 可以看出，5 台风机根据使用目的不同，被分为训练集，测试集和最终提交集。其中，风机 15 与风机 21 为训练集，数据中含有状态标签，包括正常、无效与结冰状态，无效为正常与结冰之间难以判断状态的时间。风机 08 为测试集，为防止根据时间匹配来预测测试集结冰状态，测试集的时间戳用数字进行了代替。测试集的状态标签未公布，每天可上传一次测试集的预测结果来评估算法的性能。风机 10 与风机 14 是最终提交数据集，与测试集相同的是时间戳被数字代替；与测试集不同的是表征时间连续性的“group”字段被删除，随机长度的叶片严重结冰数据也被删除，也仅有两次提交预测结果的机

会，两次预测结果的最好成绩作为复赛的最终成绩。

由以上分析可知，本次比赛的数据主要有以下三个特点:

（1）监测数据数量大、时间久、变量多。监测数据共包括5 台风机，2 个月，一百多万个时间戳的 28 个连续数值型监测变量。28 个变量包括时间戳、风速、发电机转速、网侧有功功率、对风角、偏航位置与速度、x 与 y 方向振动加速度、环境温度、机舱温度、叶片变桨角度与速度、变桨电机温度、变桨电机开关温度与开关充电器直流电流等。因为每台风机有三个叶片，因此与叶片变桨相关的变量均出现了三次，这些变量提供了相似的信息，如果将其一起输入分类器，易造成算法对训练数据的过拟合。

（2）监测数据的状态标签分布严重非平衡。训练集状态分布如表 3-5 所示，正常状态标签的时间戳约占全部监测数据的90%。非平衡的数据易造成算法对少数类识别率低，而本次比赛的风机叶片结冰预测问题恰为对少数类的识别问题。

表 3-5:　训练集状态标签分布

风机编号	总时间戳数	正常时间戳数	故障时间戳数	无效时间戳数
15	393836	350255(88.9%)	23892(6.1%)	19739(5.0%)
21	190494	168930(88.7%)	10638(5.6%)	10926(5.7%)

（3）监测变量中，风速、发动机转速、网侧有功功率、温度等变量经过标准化方法处理，失去原始的物理意义。以风速变量为例，风机 15 的风速监测曲线图如图 3-26 所示，风速变量分布在-2 到 6 之间，而风速本身作为标量无负值。这些标准

化的数据不利于理解风机的工作过程。

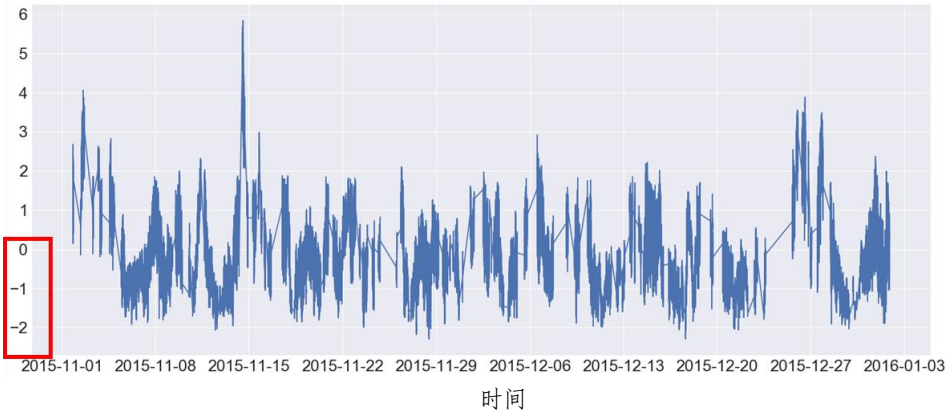


图 3-26：风机 15 风速监测曲线图

4. 方法

（1）方法分析流程

本次比赛本小组的方案为基于敏感特征的逻辑回归不平衡预测算法。主要流程如图 3-27 所示，方案分为离线训练和在线预测两部分。离线训练包括数据预处理、敏感特征选择与构建、逻辑回归分类器构建、最优阈值选择 4 步；在线预测同样要经过数据预处理，提取选择和构建好的特征，将其输入训练好的逻辑回归分类器，并将逻辑回归输出与最优阈值比较，得到分类结果。

a)数据预处理

数据预处理分为两步：数据分组与数据筛选。由背景调研和文献简介可知，风机叶片结冰会影响风机的发电功率，因此数据预处理考虑将风速与发电功率的关系显性化。显性化过程如图 3-28 所示。数据分组是将数据按照“group”将数据分组，对

于训练集和测试集，每组约含有 100 个时间戳，以组内均值作为该组特征；而对于最终提交集，数据缺少“group”字段，直接将每 100 个时间戳划分为一个组，同样以组内均值作为该组特征。该处理是因为风机叶片转动惯性减少了风机发电瞬时功率与瞬时风速之间的相关性，取组内均值则可降低惯性对其的影响。数据筛选则是通过变桨速度，筛选未达满功率发电的数据，进行后续结冰监测判别。因为从训练数据中看出，当风机满功率发电时，无结冰事件。而数据筛选截取了平均功率随平均风速近似线性变化的部分。

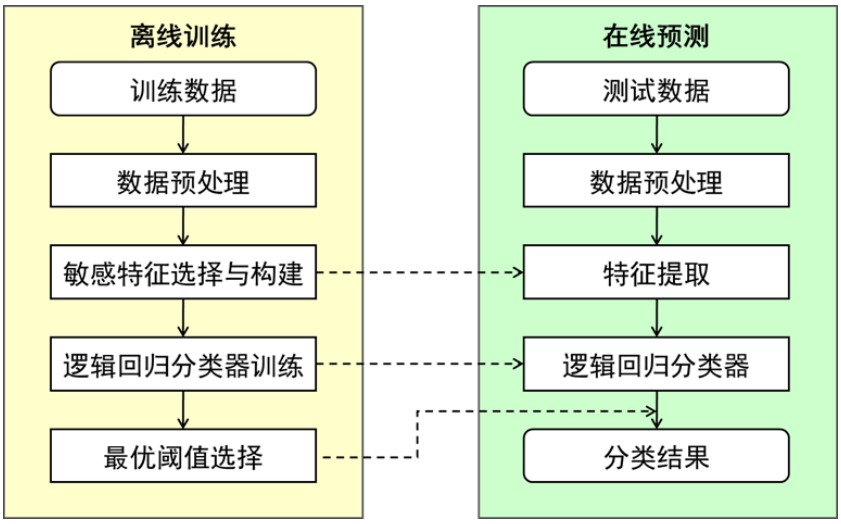


图 3-27：方案流程图

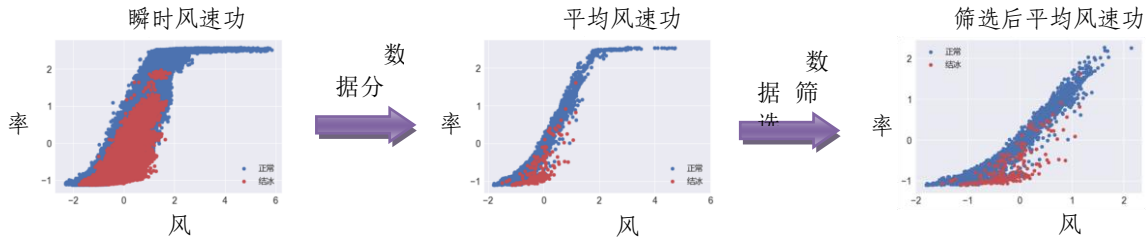


图 3-28：数据预处理

b) 特征选择及构建

经过数据预处理后将选择和构建风机叶片结冰预测的敏感表征特征。首先，分析各原始特征之间的相互关系。如图 3-29 所示，除了风速与功率之外，环境温度与机舱温度也对风机叶片结冰较为敏感。尝试分析其原因，可能是因为叶片结冰影响了风机风速与功率之间的关系，导致机舱内主传动链发热增加，使机舱温度处在较高水平，机舱内外温度差较低。

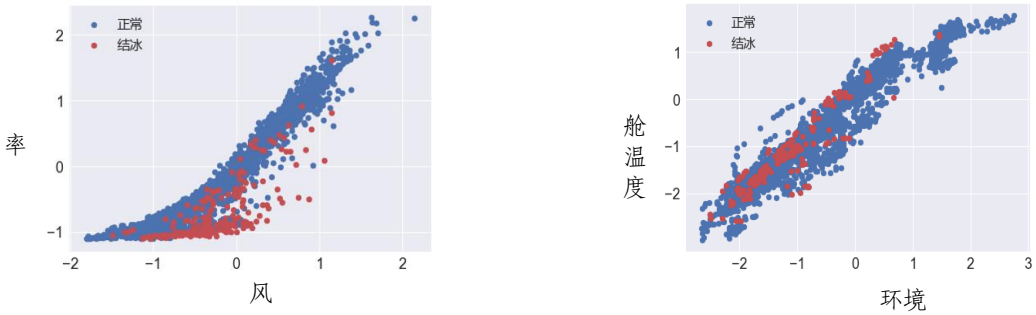


图 3-29：两组敏感特征关系图

在选择两组敏感特征之外，还提取了对结冰极为敏感的风速与功率非主成分方向投影特征，该特征利用主成分分析技术对风速与功率变量进行主成分分析。如图 3-30 左图所示，黑色

箭头方向为主成分方向，代表风速对功率的影响关系；而非主成分方向即为绿色箭头方向，可一定程度的表达结冰故障情况。构建的特征即为风速与功率在非主成分方向的投影，特征分布如图 3-30 右图所示，可以看出，正常状态和结冰状态存在明显的划分。

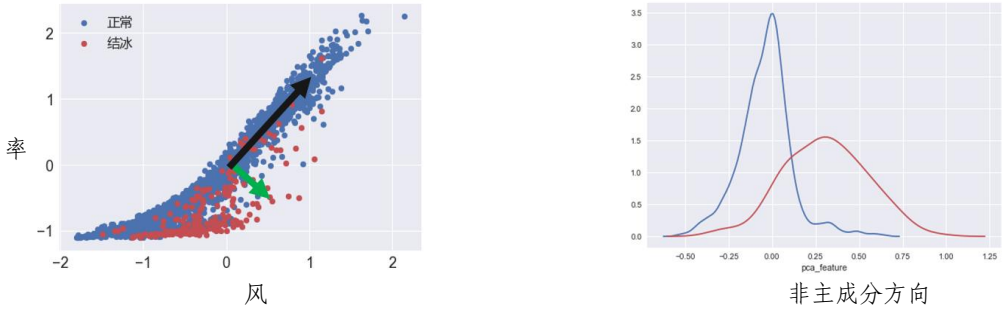


图 3-30：风速与功率非主成分方向投影特征

c)分类器构建及最优阈值选择

为使风机叶片结冰预测自动化和智能化，在提取敏感特征后将其输入机器学习模型自动判别风机的结冰状态。本方案采用的机器学习模型为逻辑回归分类器，逻辑回归函数表达式为：

$$h(\boldsymbol{x})=\frac{1}{1+e^{-\boldsymbol{\omega}^T\boldsymbol{x}+b}} \tag{1}$$

式中 $\boldsymbol{x}\in R^{n\times 1}$，表示输入特征，$n$ 为输入样本维数，$\boldsymbol{\omega}\in R^{n\times 1}$，$b\in R$，表示逻辑回归函数的参数。从概率统计角度简单解释(非严格推导)逻辑回归函数，当正反两类样本的特征均服从高斯分布时，即：

$$p(\boldsymbol{x}\mid y=0)\sim N(\boldsymbol{\mu_0},\boldsymbol{\Sigma_0}),\quad p(\boldsymbol{x}\mid y=1)\sim N(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1}) \tag{2}$$

式中，$N(\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})$和$N(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})$为高斯分布，其中$\boldsymbol{\mu_0}$和$\boldsymbol{\mu_1}$为均值，$\boldsymbol{\Sigma_0}$和$\boldsymbol{\Sigma_1}$为标准差。逻辑回归函数的输出可表示为预测为正类的后验概率，贝叶斯公式展开:

$$h(\boldsymbol{x}) = p(y=1|\boldsymbol{x}) = \frac{p(y=1)p(\boldsymbol{x}|y=1)}{p(y=1)p(\boldsymbol{x}|y=1)+p(y=0)p(\boldsymbol{x}|y=0)} \tag{3}$$

在方案中，我们希望通过两类样本的特征的先验概率对其进行分类，因此分类阈值应取为:

$$p(\boldsymbol{x}|y=1) = p(\boldsymbol{x}|y=0) \tag{4}$$

那么，当正反两类样本为平衡数据时，即:

$$p(y=0) = p(y=1) = 0.5 \tag{5}$$

将（4）式与（5）式带入（3）式，可计算逻辑回归函数的分类阈值应为 0.5。然而，当正反两类样本为不平衡数据时，假设正类出现的概率为 0.1，即:

$$p(y=0) = 0.9, \quad p(y=1) = 0.1 \tag{6}$$

将（4）式与（6）式带入（3）式，同理可计算逻辑回归函数的分类阈值应为 0.1。

从上述分析可知，当正反两类样本为不平衡数据，需要调整逻辑回归的分类判别阈值，以适应数据的不平衡性。

（2）方法的独到之处

1）算法简单，响应速度快，易于工程应用;

2）通过对风机运行原理及结冰对监测数据影响的分析，选择了四个原始特征并提取了对风机叶片结冰预测较为敏感的风速与功率非主成分方向投影特征;

3）为处理具有严重不平衡性的特征数据，提出通过改变分类阈值来使逻辑回归分类算法适合非平衡数据的方法。

5. 验证

（1）对该数据的验证

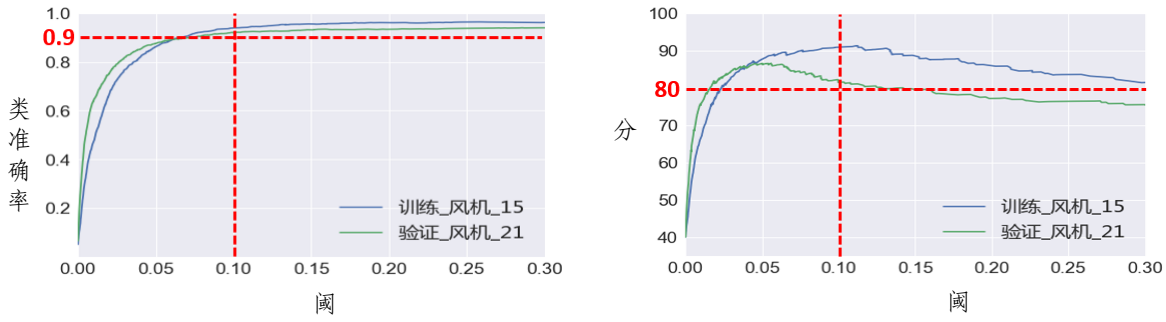


图 3-31：实验一结果

通过对训练数据进行交叉验证，选择最优的分类阈值。评估函数为分类准确率与比赛的得分函数，分类准确率评估多数类的识别情况，比赛的得分函数评估对少数类的识别情况。实验一的训练数据为风机 15，验证数据为风机 21，实验结果如图 3-31 所示；实验二的训练数据为风机 21，验证数据为风机 15，实验结果如图 3-32 所示。

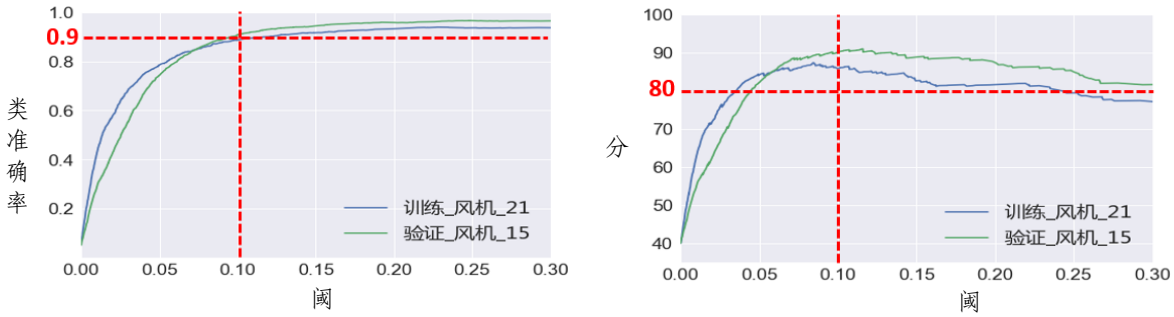


图 3-32：实验二结果

从实验结果中可以看出，随着分类阈值的升高，分类准确率即对多数类的识别能力不断增加，且在初始阶段变化明显，随后变化逐渐平缓，当阈值取为 0.1 时，分类准确率均已超过 90%；而比赛得分函数则不像分类准确率一样单调变化，得分先随着阈值的升高而升高，当阈值取为 0.1 左右时，得分达到最高点，之后随着阈值的升高而下降，当阈值取为 0.1 时，所有得分均超过 80 分。综上所述，0.1 为合理的分类阈值。

最终的实验方案如表 3-6 所示，预测结果及得分如图 3-33 所示。图中以 5 台风机的风速变量表示时间，训练集中红色区域表示实际风机叶片结冰时间，预测集和最终提交集红色区域表示预测风机叶片结冰时间。从图中可以看出，对应关系比较明显，预测较为准确。

表 3-6：最终实验方案

特征选择	风速，网测有功功率，环境温度，机舱温度，风速&功率非主成分方向投影特征

训练数据	风机 21 与风机 15 的全部数据
测试数据	风机 08、风机 10 与风机 14 的全部数据
分类算法	逻辑回归
判别阈值	0.1

如图 3-33 所示，本方案在训练集的得分为 86 分，测试集的得分为 85 分，最终提交集的得分为 75 分。在八百多支参赛队伍中最终取得第五名。

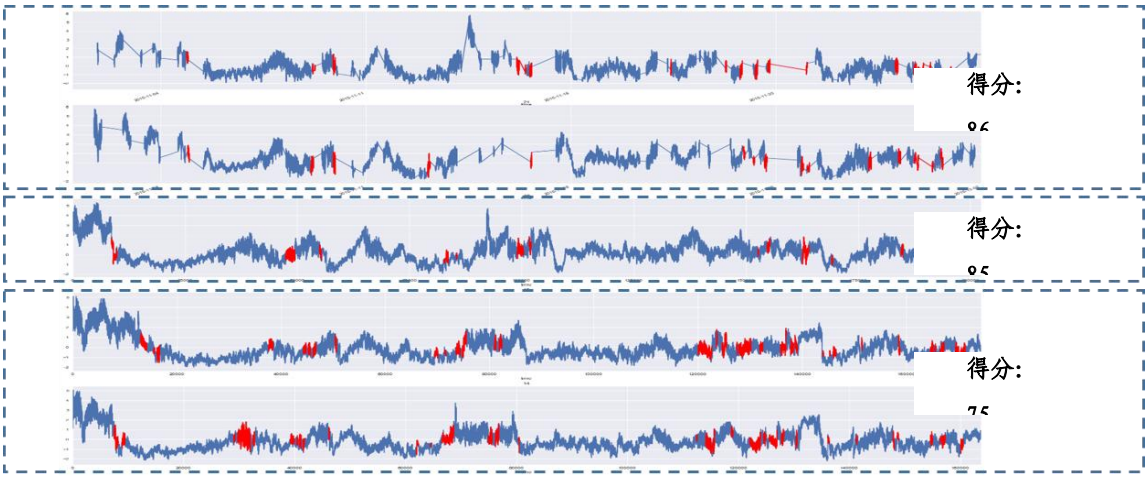


图 3-33：最终预测结果及得分

（2）与其他方案的对比

除采用手动提取特征加逻辑回归分类算法，本团队还尝试采用了深度学习框架，卷积神经网络与 LSTM 循环神经网络，从原始数据出发，自动构建特征进行分类。尽管最后学习得到的对数损失小于手动提取特征加逻辑回归分类算法，但是比赛得分却远达不到后者的表现。

6.结果分析与经验总结

（1）经验教训

本方案从工程角度出发，分析了风机叶片结冰现象对风机发电产生的影响，并以此为切入点，进行数据分组和数据筛选的预处理，构建对风机叶片结冰极为敏感的风速与功率非主成分方向投影特征，改变逻辑回归的分类阈值使其适用于不平衡分类。方案简单易用，并且对多数类和少数类均有较好的识别能力。

但是本方案的数据分组过程即为对原始数据的降采样过程，丢失了很多信息，导致算法的分类精度收到限制。

（2）计算速度与使用的硬件

本小组使用的硬件为台式机，处理器为 IntelI CoreI i5-4460，安装内存 8G，操作系统为 Windows7 64 位，编程语言为 Python，集成开发环境为 Spyder，主要使用的工具包有 Numpy，Pandas，Matplotlib，Scikit-learn。算法的训练速度为 0.003 秒，测试速度为 0.001 秒。

（3）可改进的地方

对没有“group”字段的最终提交集进行数据分组时，每 100 个时间戳硬划分为一组，有些组之中可能不具有时间连续性。根据对测试风机 08 的得分评估，硬划分的结果比按照“group”字段划分的结果低 1.5 分。可根据叶片变桨角度等变量寻找时间连续性，对数据进行分组处理。

7. 其他一些见解

低温恶劣环境给叶片乃至整个风机都带来了严峻的考验，尤其是低温等因素所导致的结冰问题，严重影响了风机的安全有序运行，因此需要对风机叶片结冰进行监测与预测。但是风机叶片结冰是一个复杂的过程，受众多环境变量的影响，因此建立严格的风机叶片结冰预测数学物理模型非常困难。而完全基于数据驱动的方法难以从纷繁复杂的原始数据中准确定位风机叶片结冰的敏感特征源，大量的原始数据处理和复杂的网络识别模型造成计算量急剧增大，不利于工业现场实用化。因此，在充分理解风力发电机的运行原理和风机叶片结冰过程的基础上，挖掘风机叶片结冰敏感特征，综合机器学习强大的特征表征和非线性映射能力，研究数模联动的技术将是未来实现风机叶片结冰预测工业现场应用的可行途径之一。

8. 参考文献

[1] http://www.cnrec.org.cn/cbw/fn/2014-12-29-459.html

[2] http://news.bjx.com.cn/html/20160401/721882.shtml

[3] Davis N N, Pinson P, Hahmann A N, et al. Identifying and characterizing the impact of turbine icing on wind farm power generation[J]. Wind Energy, 2016, 19(8):1503-1518.

[4] Shi Q. Model-based Detection for Ice on Wind Turbine Blades[D]. NTNU, 2017.

[5] Muñoz C Q G, Márquez F P G, Tomás J M S. Ice Detection Using Thermal Infrared Radiometry on Wind Turbine Blades[J]. Measurement, 2016, 93:157-163.

[6] Berbyuk V, Bo P, Möller J. Towards early ice detection on wind turbine blades using acoustic waves[M]. International Society for Optics and Photonics, 2014.

[7] Kim D G, Umesh S, Song M, et al. A fiber-optic ice detection system for large-scale wind turbine blades[C]// Optical Modeling and Performance Predictions IX. 2017:10.

[8] Khaliullin E. Wind turbine indirect ice detection and operational analysis[J]. 2016.

[9] Davis N N, Byrkjedal Ø, Hahmann A N, et al. Ice detection on wind turbines using the observed power curve[J]. Wind Energy, 2016, 19(6): 999-1010.

[10] Cattin R, Heikkilä D U. Evaluation of ice detection systems for wind turbines[J]. Meteotest, Bern, 2016.

[11] 梁颖, 方瑞明. 基于 SCADA 和支持向量回归的风电机组状态在线评估方法[J]. 电力系统自动化, 2013, 37(14):7-12.

[12] Zhang Z Y, Wang K S. Wind turbine fault detection based on SCADA data analysis using ANN[J]. Advances in Manufacturing, 2014, 2(1): 70-78.

[13] 董玉亮, 李亚琼, 曹海斌,等. 基于运行工况辨识的风电机组健康状态实时评价方法[J]. 中国电机工程学报, 2013, 33(11):88-95.

[14] 黄元维. 基于支持向量机的风电机组主轴轴承故障诊断[J]. 仪器仪表用户, 2016, 23(11): 88-92.

[15] https://www.pruftechnik.com/adwords-lp/condition-monitoring-online-systems.html

[16] http://www.skf.com/portal/skf/home/products?contentId=265076&lang=en

[17] http://www.vipzhuanli.com/pat/books/201310603324.X/2.html

[18] Skrimpas G A, Kleani K, Mijatovic N, et al. Detection of icing on wind turbine blades by means of vibration and power curve analysis[J]. Wind Energy, 2016, 19(10): 1819-1832.

[19] 2016 年中国风电装机容量简报. 中国可再生能源学会风能专业委员会, 2017.

四、方法论总结

工业大数据不同于互联网大数据的特点，决定了其处理方法的不同。中国工程院院士孙家广与清华大学软件学院院长王建民也曾经在《大数据系列报告之一：工业大数据白皮书》指出，**工业大数据具有一般大数据的特征，即海量性与多样性等，但在此基础上也有其独有的价值性、实时性、准确性、与闭环性。**美国国家科学基金会智能维护系统中心主任李杰教授所著的《工业大数据》一书中论述了工业大数据与互联网大数据不同：**在数据量的需求上，互联网大数据往往需要大量样本的数据，而工业大数据需要的与其说是量的多，不如说是工况的全；工业大数据通常对数据的质量的要求也比互联网大数据更高；对数据属性意义的解读，工业大数据分析更加强调特征之间的机理性关联。**

针对本次竞赛的数据而言，其具体特点体现在：

一是数据的量并不大。本次竞赛的数据量无论是从机器数量、参数个数、还是积累时间来讲，数据量均不大。但是，这在工业数据分析中是很常见的现象。由于工业信息化起步比互联网晚，而且并非所有企业都有存储、整理数据的早期规划，所以可供训练数据驱动模型的历史数据量并不多。然而，这并不代表工业中产生的数据不多。由于工业设备自动化的不断进步，其产生的数据仍然具有高速、大量、来源多样的大数据特点。存量数据少、新产生数据庞大，是处理工业数据时会遇到

的典型场景。这就决定了，在本次工业大数据创新竞赛中，对于数据量要求非常高的机器学习算法（data-hungry algorithms）不是最优的选择。

二是数据质量不高。在本次竞赛中，数据质量的问题主要体现在健康与故障状态数据的不平衡。工业设备往往造价高昂且可靠性高，故障发生后的运行时间不会太长，用户做破坏性试验的成本也很高，导致故障数据与健康状态数据相比经常占很少的比例。面对这样的问题，要求在建模过程的数据预处理中，通过重采样、欠采样、或者其他方法来平衡健康与故障状态的数据，从而增强模型的准确度以及泛化能力。

三是领域知识要求高。风机结冰故障会造成哪些现象、以及由哪些原因导致这种结冰故障，均需要深厚的风机行业知识才能了解。进行数据驱动的预测性建模时，这种领域知识将对风机结冰的预测起到事半功倍的作用。对于这类挑战，获取必要的领域知识通常是更加可靠的选择。加入基于机理的数据预处理、特征计算，都会在不增加计算成本的情况下，增强模型的泛化能力。

对于这三方面的挑战，各个团队采用了不同的建模策略。通过本次竞赛解法的分析，这些解法主要可以被分为两种：完全基于机器学习算法的模型，和与机理融合的模型。

结合此次竞赛结果可以发现，领域知识的融入往往可以增强模型的泛化能力与准确性，其结果通常会优于没有领域知识

的纯粹机器学习模型，尤其在数据质量不好的情况下。在增加机器学习模型复杂度后，其准确性有可能与“机理+简单模型”的结果相当。但是，其代价是模型训练时间随着其准确度的上升而显著增加，并且复杂性增加后，模型本身变成黑盒子，可解读性变差。此外，通过分析数据建模方法不难发现，虽然宏观的流程基本相似，但对各个环节的不同处理方式，决定了模型的性能与效果。同时，数据的预处理对最后的预测结果影响不亚于模型本身，在工业大数据预测性建模中也要受到足够的重视。