

文章编号: 2096-1618(2018)06-0632-07

基于随机森林算法的电线覆冰检测技术

罗扬焱, 卢会国, 蒋娟萍, 曼世超

(成都信息工程大学电子工程学院, 四川 成都 610225)

摘要:随机森林是21世纪提出的基于分类树的算法,具有学习过程快速、运算速度快、稳定性好、预测精度高的优点。使用2017年8月雅安市泥巴山站点的电线覆冰数据训练一个用于预测覆冰现象是否出现的预测模型,并使用2017年11月–2018年1月的观冰数据对这一模型进行测试。测试结果表明该模型具有92.16%的查准率和88.26%的查全率。而使用2017年11月–2018年1月的电线覆冰数据训练得到的覆冰重量回归模型较线性回归模型有着更小的均方误差,随机森林回归模型的非线性特性能够更好地拟合覆冰重量。实验表明:随机森林算法既可以用于预测覆冰现象的出现,也可用于预测覆冰重量的变化;即随机森林算法在电线覆冰检测领域具有巨大潜力。

关键词:随机森林;电线覆冰检测;查准率;查全率;均方误差

中图分类号:TP181

文献标志码:A

doi:10.16836/j.cnki.jcuit.2018.06.006

0 引言

雨淞、雾凇凝附在导线上或湿雪冻结在导线上的现象,称为电线覆冰。电线覆冰是一种气象灾害,对电力系统的正常运行有非常严重影响。覆冰可能导致电线舞动、杆塔倾斜、倒塌、断线,从而引发通信不畅、停电断水等问题。2008年冬季,中国出现了历史罕见的冰冻天气,致使13个省电力系统运行受影响,最严重地区的电网近乎瘫痪,此次冰冻灾害造成了难以计数的经济损失^[1]。因此准确预报电线覆冰,对指导输电线路上的冰冻灾害防治具有重要意义。

国内外的研究人员对电线覆冰的形成条件和形成机理进行了详细的研究和论述。牛生杰等^[2-3]对电线积冰的微物理机制和气象影响方面进行了广泛研究;刘雪静等^[4]在两次高压电线积冰的数据基础上,分析了电线积冰期间的气象条件。众多的研究成果使对覆冰过程有一个比较完善的认识,以此为基础,可以由不同的研究方法产生出不同的覆冰厚度预测模型。影响覆冰的气象因素有温度、降水、湿度、风速和风向等。除此以外,导线本身的悬挂高度、线径粗细、电场及电荷的分布等也会对覆冰的厚度、质量产生影响。迄今为止,提出的覆冰预测模型大致可以分为经典覆冰预测模型和统计覆冰预测模型。

经典覆冰预测模型是通过热力学、流体力学等方面的物理知识来建立守恒方程,从微观层面将覆冰的

重量同各类气象要素建立联系,主要代表有 Stallabrass 公式^[5]和 Finstad 公式^[6]等。这些经典覆冰预测模型已经能够获得较理想的实验结果,然而经典覆冰预测模型中涉及众多的物理参数,如液态水含量、液滴半径谱、液滴撞击速度之类的参数在自然环境下很难做到精确测量,并且,经典覆冰预测模型或多或少都存在着对实际物理过程的简化,并不能全面地模拟覆冰形成的物理过程。这些问题成了限制经典覆冰预测模型提升准确率的主要因素。

随着大数据时代的到来,发展了一系列基于机器学习的统计覆冰预测模型,相比于基于物理知识建立起来的经典覆冰预测模型,统计覆冰预测模型不关注覆冰形成的物理过程,通过对覆冰观测资料进行统计分析,从宏观层面将气象要素与覆冰重量联系起来。早期的统计预测模型所使用的是多元回归方法^[7],但是电线覆冰本质上是复杂的非线性过程,因此利用线性拟合的手段难以对覆冰重量进行精确拟合。针对电线覆冰过程的非线性特点,近年来随着机器学习领域的飞速发展,提出了许多能够解决非线性问题的覆冰预测模型。郑振华等^[8]提出了一种遗传算法与BP神经网络相结合的预测模型。黄宵宁等^[9]提出了一种基于数据驱动算法和最小二乘支持向量机的覆冰预测模型。尹子任等^[10]提出了一种基于粒子群算法优化支持向量机的电线覆冰预测模型。然而,人工神经网络与支持向量机存在着调参过程复杂、训练时间长、容易陷入局部极小值。参数依靠经验设置,缺乏理论指导。

随机森林是一个以随机方式建立,包含多个决策树

收稿日期:2018-03-03

基金项目:四川省教育厅重点科技计划资助项目(14ZA0170)

的分类器。其输出的类别是由各个决策树输出的类别共同决定的。模型结构易于人类理解,训练完成的分类器能够显示出数据中各类特征的重要程度。相较于神经网络和支持向量机,随机森林算法具有训练时间短、泛化能力强、不易过拟合、训练参数少、易于实现的特点。相较于一般回归分析,随机森林能够自动进行特征选择,并且具有更好的非线性处理能力。因此,使用随机森林算法可以同时处理分类与回归问题。在分类问题上,随机森林算法可以得到一个结构简单、易于理解、训练成本低、泛化能力强的分类模型。而在回归问题上,随机森林算法也可用于处理非线性回归问题。

从理论上讲,统计覆冰预测模型其精确程度低于经典覆冰预测模型,但在由于其具有所需气象参数数量少、参数容易获得等特点,在实际的覆冰预测业务中具有巨大的潜在价值。而且,训练数据的数量以及其多样性对统计预测模型的性能有着重要影响,随着气象数据的不断积累,观冰站数量的不断增加,各种新式机器学习算法的不断改进,统计预测模型的预测准确性还有很大的提升潜力。

1 算法介绍

随机森林算法是在决策树算法的基础之上所发展起来的一种算法。决策树是一个树结构,其每个非叶节点表示针对某一个特征属性的判定环节,其分支表示该特征属性在判定之后的输出,而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点出发,在非叶节点上对相应的特征属性进行判定,并按照判定结果选择分支路线,直到到达存放结果类别的叶节点。生成决策树的核心问题就是寻找最佳划分特征来确定每一个非叶节点的判定条件。经典决策树算法中的C4.5^[11]算法使用信息增益率作为评价标准具有最大信息增益率的特征将被认为是最佳划分特征,如此便能得到一个非叶节点和多个次级数据集。重复进行上述操作,直到再无特征可划分或达到最大深度,最后按照少数服从多数的原则来确定叶节点的类别。CART^[12]算法与C4.5算法流程上一致,唯一区别在于CART算法使用基尼指数作为评价标准。需要说明的是,C4.5算法构造的决策树只能用于分类问题,CART算法构造的决策树既能处理分类问题也可用于回归问题。

生长完成的决策树存在着过拟合问题,为获得更好的泛化能力,必须要对决策树进行剪枝,即删去一些可能会招致过拟合的节点。常用的方法是使用测试数

据集,通过比较某一节点在删除前后正确率是否获得提升来判断该节点是否需要剪枝。

2001年Breiman提出了使用多个决策树来进行投票决策的随机森林算法^[13]。随机森林算法以随机抽取的方式从整个训练数据集中抽取出一个子集用于生成一棵决策树,并且在每次生成非叶节点时,都是在随机选取的多个特征中寻找最佳划分特征。多次重复上述步骤就可以获得多个形态不同的决策树,也就是一个决策森林。在得到森林之后,如果需要对一个新的数据样本进行测试,就让森林中的每一棵决策树分别对这个数据样本进行判断。最后统计所有决策树的判定结果,从而确定最终的判定结果。

决策树生成的随机性以及森林决策的组合性使得随机森林算法相比较于单棵决策树具有更好的泛化能力,可以有效避免过拟合问题。因此随机森林算法一般不需要剪枝过程。

2 实验描述

2.1 数据准备

实验所用的数据来源于成都信息工程大学所研发的电线覆冰自动观测系统,系统被安装于雅安市汉源县泥巴山站点(29°38'N, 102°36'E, 海拔约2450 m)。系统基于称重法原理能够完成对覆冰重量的观测,并且能够对温度、气压、风速和风向进行实时观测。观测到的重量数据不仅包含了覆冰重量,还包括导线重量以及风和其他因素所造成的干扰量。

整个数据集包含了从2017年7月-2018年1月绝大部分的观测数据。实验将使用东西合重、南北合重、温度、风速4种数据类型。实验包括两个部分:第一部分使用温度、风速以及覆冰标注作为训练数据,期望训练出一个分类模型,其能够根据输入的温度和风速条件预测出是否会出现覆冰;第二部分使用温度和风速作为训练数据(以下简称RFC),期望训练出一个回归模型(以下简称RFR),其可以根据输入的温度和风速预测出覆冰重量。因为实验即要解决分类问题也要解决回归问题,所以实验中的森林将会使用CART算法来生成决策树。

以随机抽取的方式随机抽取出可能出现覆冰记录的2017年11月-2018年1月的观测数据中的10%作为分类模型和回归模型的测试数据,将剩下的数据作为两种模型的训练数据。随机森林算法在做分类预测时,需要预先对训练数据进行分类标注。而实验所使

用的数据均来自于电线覆冰自动观测系统,该观冰站属于无人站,因而没有人工的覆冰标注。对此,实验以东西合重与南北合重两个称重数据为目标,假设称重数据过大表示此时出现了覆冰现象。实验将首先使用 2017 年 8 月的 4300 条观测数据训练出一个基于孤立森林算法^[14]的异常检测分类模型,再使用此分类模型对 2017 年 11 月–2018 年 1 月的 85000 条观测数据进行覆冰标注。同时,其标注结果也将被假设为数据是否出现覆冰的真值。

基于孤立森林的异常检测分类模型(以下简称 IF)使用东西合重与南北合重两个数据类型作为训练数据,训练完成的模型将会根据输入数据的东西合重南北合重判断出其是否处于覆冰状态。

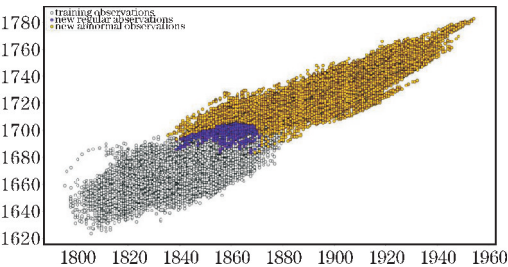


图 1 训练数据的标注

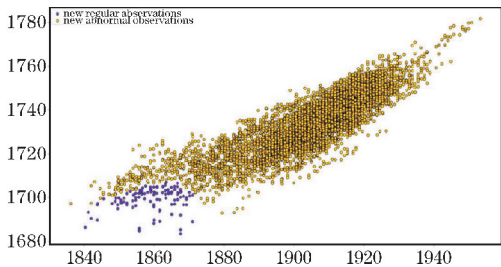


图 2 测试数据的标注

以 2017 年 8 月的观测数据为基准,IF 分类模型会把远离标准数据点的数据点归类为异常数据点,可以将这些数据点认为是出现覆冰现象的数据点。

2.2 评价标准

由图 1、图 2 可以看出,2017 年 11 月–2018 年 1 月的数据中覆冰数据要远远多于未覆冰数据,因此对覆冰情况进行预测就变成了一个类别不平衡问题,为了使训练数据中两类数据的数目接近,实验将会采用欠采样技术,即去除部分覆冰观测数据,再使用这个处理过后的训练数据集去训练覆冰预测模型。覆冰预测模型旨在成功预测出覆冰是否会出现,从而帮助电网工作人员及时做好灾害预防工作。因此,将未覆冰的数据错误预测为覆冰数据,或是将覆冰数据错误预测为未覆冰数据,这二者之间的错误代价应该是不同的。

实验并不能简单地用正确率作为评价标准,因为正确率这个标准隐式的假设了均等错误代价。综上所述,实验将使用查准率和查全率作为评价标准。

表 1 二分类混淆矩阵

真实结果	预测结果	
	覆冰	未覆冰
覆冰	TP	FN
未覆冰	FP	TN

查准率 $P = \frac{TP}{TP + FP}$ (1)

查全率 $R = \frac{TP}{TP + FN}$ (2)

查准率的含义是,测试中被认为出现了覆冰现象的结果中有多少是真正的出现了覆冰现象。查全率的含义是,测试中真正出现了覆冰现象的数据有多少被正确预测。越高的查全率就表示该模型能够更好地预测覆冰现象的出现,越高的查准率表示模型能够更准确的区分覆冰现象是否发生。

回归问题常用的评价标准是均方误差(MSE):

$$MSE = \frac{\sum_i^n [f(i) - y_i]^2}{n}$$
 (3)

均方误差是反映估计值与真实值之间差异程度的一种度量。均方误差越大,表明预测结果与真实值的差异也越大。

由图 1 和图 2 可以看出,数据中覆冰数据的数量要远远多于未覆冰数据的数量。假设测试数据中有 99 条覆冰数据,只有 1 条未覆冰数据。那么,模型只要把测试数据全部判定为覆冰数据就能获得 99% 的查准率和 100% 的查全率。可是这种模型毫无意义,因为实际上并不能区分覆冰数据和未覆冰数据。为了回避这个陷阱,实验将对测试数据中由 IF 模型标注为覆冰状态的数据进行“欠采样”,使处理过后的测试数据中覆冰与未覆冰数据所占比例大致相当。

3 实验过程

3.1 分类模型实验

随机森林算法中对模型性能会产生重要影响的参数就是森林中决策树的数量,实验将尝试选取不同的决策树数量,并以此参数来训练模型。根据 5 次实验的平均测试结果寻找最优决策树数量。

表 2 决策树数量对 RFC 模型查准率的影响

决策树数量	第一次实验	第二次实验	第三次实验	第四次实验	第五次实验	平均查准率
10	0.891	0.903	0.892	0.871	0.886	0.8886
20	0.896	0.965	0.937	0.913	0.908	0.9238
25	0.905	0.921	0.912	0.917	0.953	0.9216
30	0.951	0.943	0.901	0.929	0.903	0.9254
35	0.921	0.941	0.918	0.948	0.886	0.9228
40	0.892	0.888	0.928	0.922	0.920	0.9100
50	0.930	0.901	0.912	0.916	0.898	0.9114
60	0.905	0.900	0.906	0.913	0.924	0.9096

表 3 决策树数量对 RFC 模型查全率的影响

决策树数量	第一次实验	第二次实验	第三次实验	第四次实验	第五次实验	平均查全率
10	0.803	0.835	0.869	0.838	0.846	0.8382
20	0.863	0.850	0.900	0.896	0.831	0.8680
25	0.905	0.854	0.920	0.855	0.879	0.8826
30	0.891	0.902	0.830	0.883	0.875	0.8762
35	0.820	0.870	0.878	0.840	0.868	0.8552
40	0.892	0.807	0.873	0.864	0.902	0.8676
50	0.846	0.827	0.798	0.831	0.851	0.8306
60	0.802	0.876	0.804	0.885	0.867	0.8468

由表 2 可知,决策树数量由 10 增加到 20 后平均查准率有一定的提升。决策树数量从 20 增加到 35 这个过程中,平均查准率没有明显变化。在这之后,决策树数量的增加反而会使平均查准率有一定程度的下降。表 3 可知,决策树数量由 10 增加到 20 后平均查全率有一定的提升,并在决策树数量达到 25 左右时达

到最佳。在此之后,决策树数量的增加反而会使得平均查全率下降。实验数据表明,决策树数量设定为 25 时,随机森林分类模型能够更好地拟合积冰数据。结合表 2、表 3 可以发现,在相同参数下,查全率几乎总是低于查准率。而且 RFC 模型在实验数据上出现漏报的概率要大于出现虚警的概率。

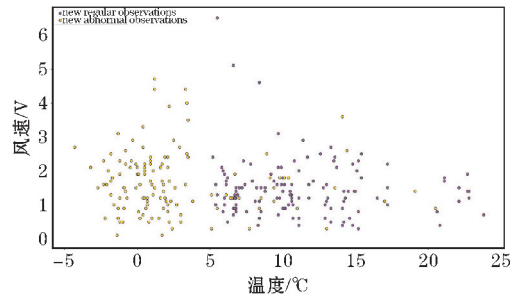


图 3 测试数据在 IF 模型中的测试结果

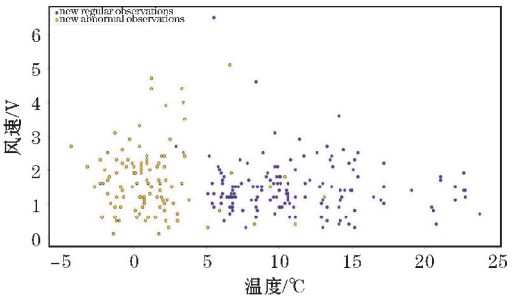


图 4 测试数据在 RFC 模型中的测试结果

图 3 所使用的 IF 模型是根据东西合重、南北合重两个称重数据得到的分类结果,图 4 的 RFC 模型则是使用温度和风速两个特征来预测是否出现覆冰。相较于图 3,图 4 中一些温度处于 5℃~20℃的数据被分类为正常数据,即没有出现覆冰现象。考虑到电线覆

冰的消融需要一定的时间,所以在较高的温度下,电线覆冰现象依然可能存在。图 3 是根据称重结果得到的判定结果,可信度更高。而 RFC 模型则是轻率地将较高温度的数据分类为正常数据,忽略了覆冰消融这一过程。因此,图 4 中不同于图 3 的分类结果很可能是

错误分类,这也就解释了实验中为何漏报概率总是大于虚警概率。电线覆冰的凝结和消融过程都需要一定时间来完成,在缺少时间信息的情况下,仅依靠其他特征来作预测,将不可避免地引入一定的误差,在对较高温数据进行预测时,这种误差尤其明显。

3.2 回归模型实验

对于覆冰重量的预测实质上是一个回归问题,随

机森林算法也可处理此类回归问题。为了获得解决回归问题的随机森林预测模型(以下简称 RFR),实验将抽取 2017 年 11 月-2018 年 1 月观冰数据中 90% 的数据,使用其中的温度、风速数据作为训练数据,并分别以东西合重、南北合重作为标注,训练出两个 RFR 模型,并使用相同的数据训练出两个线性回归模型用作对比。观冰站剩余的 10% 数据将用于对两种预测模型进行测试。

表 4 决策树数量对东西合重回归值的影响 g

决策树数量	第一次实验	第二次实验	第三次实验	第四次实验	第五次实验	平均
10	109.023	109.212	110.675	109.542	109.757	109.6418
20	112.576	114.679	109.952	111.43	112.264	112.1802
30	109.729	111.738	112.946	109.597	111.067	111.0154
50	110.349	111.951	110.917	115.382	111.3	111.9798
80	111.54	111.37	111.593	109.374	112.457	111.2668
100	107.664	110.985	110.468	113.867	112.717	111.1402
110	111.467	111.824	110.503	111.74	111.392	111.3852
120	107.829	107.083	106.715	109.55	112.267	108.6888
130	111.517	106.853	112.118	107.86	110.359	109.7414
140	109.899	110.374	109.205	109.118	110.849	109.889
150	108.579	107.839	108.592	112.92	110.777	109.7414

表 5 决策树数量对南北合重回归值的影响 g

决策树数量	第一次实验	第二次实验	第三次实验	第四次实验	第五次实验	平均
10	107.902	109.794	109.721	108.71	108.59	108.9434
20	111.498	111.914	109.507	108.603	112.066	110.7176
30	109.152	109.832	109.588	107.162	111.567	109.4602
50	109.813	110.744	108.304	112.843	109.559	110.2526
80	109.857	109.46	109.688	109.214	109.703	109.5844
100	106.106	106.534	107.233	112.059	109.117	108.2098
110	111.261	110.404	109.898	112.8	112.478	111.3682
120	108.952	107.207	106.06	108.612	111.307	108.4276
130	110.233	109.877	109.696	107.058	107.741	108.921
140	108.29	110.116	109.733	108.357	111.212	109.5416
150	109.149	109.06	108.863	111.946	112.592	110.322

从表 4、表 5 可以看出,数次测试结果并无明显变化,决策树数量的改变没有对模型性能产生明显的变化。Breiman 曾指出,随机森林算法随着决策树数量的增加,其测试误差将会收敛到一个界限,算法性能不

会再发生明显变化。由表中数据可知在决策树数量取 10 时,RFR 模型便可以很好的拟合实验所使用的结冰数据。东西合重最佳均方误差为 109.6418 g,南北合重最佳均方误差为 108.9434 g。

表 6 线性回归模型的测试结果 g

	第一次实验	第二次实验	第三次实验	第四次实验	第五次实验	平均
东西合重	112.323	124.071	123.677	122.006	124.314	121.2782
南北合重	112.286	113.545	113.801	112.192	114.813	113.3274

由表 6 知,线性回归模型对于东西合重的均分误差为 121.2782 g,对于南北合重的均方误差为 113.3274 g。两个线性回归模型的均方误差不仅大于 RFR 模型在决策树数量为 10 时所获得的均方误差,而且大于表 4 和表 5 中所有测试结果的均方误差值。

在东西合重上,两个模型的差距为 11.6364 g,在南北合重上的差距为 4.384 g。随机森林回归模型比线性回归模型,特别是在对东西合重的回归预测上,能够更好地拟合数据。

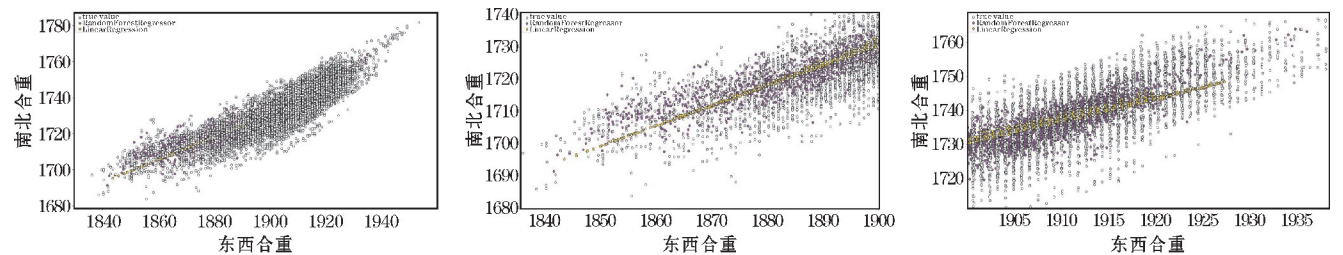


图 5 两种模型回归值与真实值的比较

图 5 画出了两种回归模型的覆冰回归数值以及测试数据的真实覆冰数值,从中可以看出,实验中真实的覆冰重量分布接近于一种非线性的带状分布。不同于人工观测所获得的观冰数据,自动观冰站获得的称重数据会受风的干扰。东西合重的某一数值上有多个南北合重数值与之对应,且南北合重各数值间的最大差距甚至能够达到 40 g,这给回归预测带来了巨大的困难。表 4、表 5、表 6 的均方误差数值均达到了 100 以上,体现出数据质量不利于做回归预测。图 5 中紫色数据点更加接近真实数据的带状分布,这也表明随机森林算法的非线性特性能够更好地拟合覆冰重量的分布。

性能产生可观的提升。电线覆冰是一个逐渐变化的过程,若要更好的拟合这一变化过程,未来的机器学习算法中应该加入时序信息。

4 结束语

参考文献:

随机森林算法能够用于处理电线覆冰问题,业务人员可以使用历史数据对预测模型进行训练,并使用预测模型对实时的气象监控数据进行分析从而预测出其是否意味着将会出现覆冰现象,覆冰重量可能会达到哪种程度。这些预测结果可以帮助业务人员对即将到来的覆冰事件进行灾害评估,从而做出更加合适的应对方案。

[1] 陆佳政,蒋正龙,雷红才,等. 湖南电网 2008 年冰灾事故分析[J]. 电力系统自动化,2008,32(11):16-19.

[2] 牛生杰,周悦,贾然,等. 电线积冰微物理机制初步研究:观测和模拟[J]. 中国科学:地球科学,2011(12):1812-1821.

[3] 牛生杰,李蕊,吕晶晶,等. 三种下垫面温度及结冰预报模型研究[J]. 地球物理学报,2011,54(4):909-917.

[4] 刘雪静,牛生杰. 两次高压电线积冰过程气象成因分析[J]. 气象科学,2016,36(2):230-235.

[5] Makkonen L. Modeling of ice accretion on wires[J]. J Clim Appl Meteor, 1984, 23(6):929-939.

[6] 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.

[7] 廖玉芳,段丽洁. 湖南电线覆冰厚度估算模型研究[J]. 大气科学学报,2010,33(4):395-400.

[8] 郑振华,刘建生. 遗传算法与 BP 神经网络相结合的输电线路覆冰厚度预测方法[J]. 电网与清

由于实验所使用的观冰数据中气象数据种类较少,并且观冰系统所使用的传感器在性能上存在一定的缺陷,导致模型使用的训练数据质量不够理性,从而影响到覆冰预测模型最终的性能,特别是回归模型的性能。使用更加丰富、更高质量的观测数据将对模型

洁能源,2014,30(4):27-30.

[9] 黄宵宁,许家浩,杨成顺,等. 基于数据驱动算法和 LS-SVM 的输电线路覆冰预测[J]. 电力系统自动化,2014,38(15):81-86.

[10] 尹子任,苏小林. 基于粒子群算法优化支持向量机的输电线路覆冰预测[J]. 电力学报,2014(1):6-9.

[11] Quinlan J R. C4. 5:programs for machine learning [J]. 1993(1).

[12] Breiman L,Friedman J H,Olshen R,et al. Classification and Regression Trees [J]. Encyclopedia of Ecology,1984,40(3):582-588.

[13] Breiman L. Random Forests [J]. Machine Learning,2001,45(1):5-32.

[14] Liu F T,Ting K M,Zhou Z H. Isolation-Based Anomaly Detection [J]. Acm Transactions on Knowledge Discovery from Data,2012,6(1):1-39.

Wire Icing Detection Technology based on Random Forest Algorithm

LUO Yang-yi, LU Hui-guo, JIANG Juan-ping, MAN Shi-chao

(College of ElectronicEngineering,Chengdu University of Information Technology,Chengdu 6100225,China)

Abstract: Random forest is an algorithm proposed in the 21st century based on classification tree. It has the advantages of fast learning process,fast operation speed,good stability and high prediction accuracy. Use wire icing datafrom Nibam-ountain site of Yaan city in August 2017 training a prediction modelused to predict whether icing phenomenon will be appeared, and use the icing data between November and January to test the model in this article. The test results show that the predictionmodel'sprecision ratio is 92.16% and its recall ratio is 88.26%. Use wire icing data between November and January train an ice weight regression model thathas a smaller mean square error rather than linear regression model. Because of the nonlinear characteristic,random forest regression model has a better ability in fitting the ice weight. The experiment shows that the random forest algorithm can be used to predict the occurrence of wire icing phenomenon ,and it also can be used to predict the change of ice weight. In other words, the random forest algorithm has greatly potential ability in the field of wire icing detection.

Keywords: random forest;wire icing detection;precision rate;recall rate;mean square error