基于随机森林算法的配网抢修故障量预测方法①

程淼海[1], 楼 俏[1], 王 琼[1], 王国军[1], 胡殿刚[2], 李韶瑜[2]

[1](国网甘肃省电力公司 兰州供电公司, 兰州 730050)
[2](国网甘肃省电力公司, 兰州 730030)

摘 要: 配网抢修是电力系统运行环节中十分重要的一环, 精益化的配网抢修管理不仅能提高电力系统的供电服务质量, 也能减少电力公司的经济损失. 本文提出一种新的配网抢修故障数量预测的方法. 首先, 基于历史数据, 以气温、风力、前一天的故障量、最大最小负荷等作为因变量, 对数据做了特征映射等预处理. 然后, 应用随机森林算法建立配网抢修故障量预测模型, 并预测不同区域、不同电网故障及非电网故障、不同电压维度下未来一天故障量. 在真实电力数据上进行了对比验证, 实验结果表明提出的方法具有较好的预测效率和准确性.

关键词: 配网抢修; 电力系统; 精益化管理; 故障量预测; 随机森林算法

Method for Fault Forecasting in Repair of Distribution Network Based on the Random Forest Algorithm

CHENG Miao-Hai[1], LOU Qiao[1], WANG Qiong[1], WANG Guo-Jun[1], HU Dian-Gang[2], LI Shao-Yu[2]

[1](State Grid Lanzhou Branch Electic Power Company of Gansu, Lanzhou 730050, China)
[2](State Grid Gansu Electic Power Company, Lanzhou 730030, China)

Abstract: Repair in the distribution network is a very important part of the power system running, the lean of the distribution network emergency management can not only improve the quality of power supply service, but also reduce the economic losses of power companies. In this paper, a new fault forecasting method of repair in the distribution network is proposed. Firstly, based on historical data, the temperature, wind, the fault of the previous day, the maximum and minimum loads, etc are regarded as dependent variables. and feature mapping and preprocessing are performed on the variables. Then, the Random Forest algorithm is applied to establish the fault prediction model of the repair of distribution network, and to forecast the future failure rate in different regions, different power grids and non-grid faults and different voltage dimensions. The experimental results on the real power data show that the proposed method has better prediction efficiency and accuracy.

Key words: repair of distribution network; electrical power system; lean management; fault forecasting; random forest algorithm

随着社会的发展, 人们的用电需求也在不断提高, 电力的供应服务质量直接影响到人们的生活质量, 然而当今的电力系统运行环节中仍存在着许多的不足. 如何提供可靠安全的供电的服务, 迅速有效地对电网故障进行抢修是电力公司急需解决的一个问题, 其中一个行之有效的方法就是配网抢修的精益化管理[1].

当今, 有众多研究者提出了许多配网抢修的策略办法等. 袁仲雄等研究者提出了基于模糊评判法的配网抢修模型[2], 金家红等研究者实现了一种配网抢修移动应用系统[3], 张敏智等研究者则是对配网抢修效率提升策略进行了探析[4]. 然而, 在众多的策略办法中鲜少看到对故障量进行预测的办法, 若能对故障量

① 基金项目:国家自然科学基金(61103175,61300104);教育部科学技术研究重点项目(212086);福建省科技创新平台建设(2009J1007);福建省自然科学基金(2013J01230);福建省高校杰出青年科学基金(JA12016);福建省高等学校新世纪优秀人才支持计划(JA13021)
收稿时间:2015-12-28;收到修改稿时间:2016-03-31 [doi:10.15888/j.cnki.csa.005357]

进行预测, 就能提前安排好抢修人员及工具等, 有利于提高抢修完成时间.

基于上述研究现状, 本文提出了一种配网抢修故障量预测的方法. 现有的配网抢修故障量预测算法有很多, 例如神经网络算法、贝叶斯算法、支持向量机算法等. 近几年来, 神经网络算法被广泛应用于电路的故障量预测, 但神经网络算法必需必须获取足够多的训练样本数量, 否则在训练时容易陷入局部最优解; 基于支持向量机的方法由于在特征多、类别结构复杂时仍有较高的分类精度, 运用也十分广泛, 但是支持向量机算法对缺失数据敏感, 在分类特征量较多时, 存在训练速度较慢、占用资源较多等问题[5]; 朴素贝叶斯算法处理不确定性问题的能力很强, 可以高效地进行多元信息的融合与表达, 因此也十分适用于故障量的预测, 但朴素贝叶斯算法除了需要假定先验分布之外, 还需要假设所有样本之间是互相独立的, 而现实世界中的数据通常难以满足独立性假设. 因此, 本文基于统计方法和数据挖掘技术, 选择应用随机森林算法[6]来建立配网抢修故障量预测模型. 随机森林由许多的决策树组成, 简单易用, 具有较好的抗噪能力, 分类错误率低, 克服了传统分类模型精度不高、容易过度拟合的问题, 预测准确率高. 因此, 研究基于随机森林算法的配网抢修故障量预测方法对提升配网抢修的精益化管理水平具有一定的参考价值.

1　随机森林算法

早在 1995 年, 贝尔实验室的 Tin Kam Ho 就结合 Bagging 方法[7]和随机子空间方法提出了随机决策森林(random decision forests). 随机森林算法是由 Leo Breiman 和 Adele Cutler 于 2001 年在决策树的基础上提出的一种组合决策树分类器的算法.

定义 1. 随机森林是一个由一组决策树分类器 $\{h(X,\ ,k)\}$ 组成的组合分类器, 其中 $\{\theta_k\}$ 是服从独立同分布的随机变量, K 表示随机森林中决策树的个数, 在给定自变量 X 的情况下, 每个决策树分类器通过投票来决定最优的分类结果[8].

随机森林的构建过程如图 1 所示, 可以分为几个步骤. 首先是利用 bootstrap 重抽样方法有放回地从原始数据样本中随机抽取 n 个与原始数据样本容量相同的训练样本; 然后利用随机方法从训练样本的 M 个特征变量中抽取 $m(m<M)$个特征变量, 来为训练样本构建决策树分类器, 所有决策树分类器就构成了随机森林; 最后对所有决策树的预测结果通过投票来得到最终的预测结果.

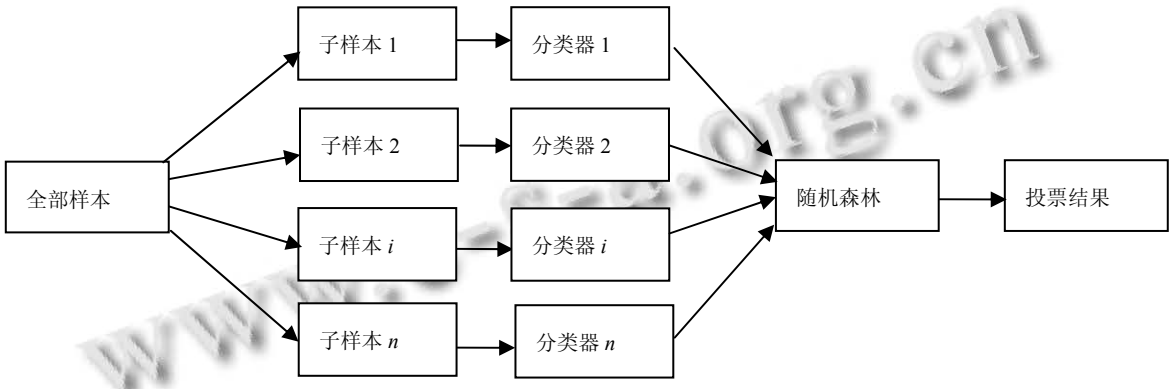


图 1　随机森林构建图

其中, 单棵决策树分类器的生成过程如图 2 所示, 首先从全部样本中随机抽取出一个训练样本, 然后等概率随机地从所有特征变量中抽取特征变量子集, 再从特征变量子集中选取最优的一个属性来进行决策树节点分裂, 最后判断样本是否训练结束, 结束则生成决策树, 否则返回继续抽取特征变量子集来进行最优属性选择. 由于各个决策树的训练是相互独立的, 因此随机森林的训练可以通过并行处理来实现, 这将大大提高生成模型的效率.

随机森林的随机性体现在了以下两个方面:

(1) 在样本抽取方面, 随机森林基于 Bagging 方法, 利用有放回重抽样的办法在原始数据样本中随机抽取和原始数据样本数量一样多的训练样本. 当样本容量很大时, 原始数据样本中会有大约 37%的样本未被抽

中. 未被抽中的样本就作为袋外数据(OBB)用来对该算法性能进行评价.

(2) 在属性选择方面, 随机森林基于随机子空间方法, 从 M 个特征变量中随机抽取 m(m<M)个特征变量, 再从这 m 个特征变量中选出一个最优特征变量来作为决策树节点分裂时的内部节点. 每棵决策树的生长都不进行剪枝.

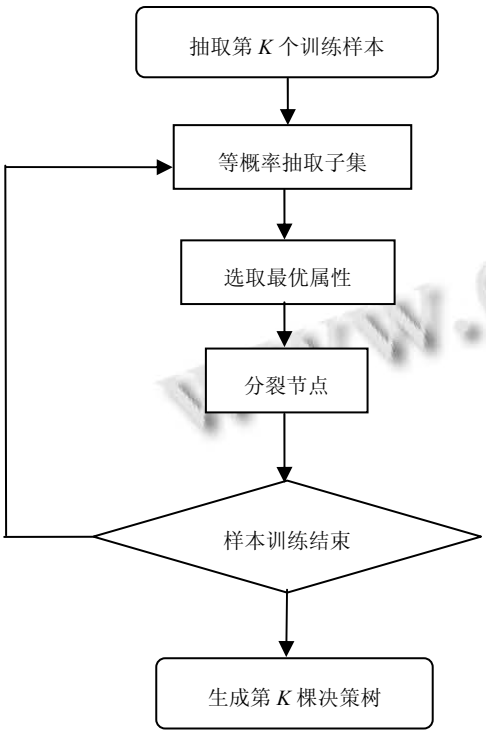


图 2 决策树生成过程

2 配网抢修故障量预测

2.1 总体处理流程

本文所研究的配网抢修故障量预测方法的总体处理流程如图 3 所示. 因为分类算法需要根据未来一天的其它特征预测故障量特征的值, 所以需要先采用预测算法预测未来一天的天气、负荷等特征的值, 然后再根据这些特征的值, 来预测未来一天的故障量. 因此, 将总体处理流程分为非故障量特征预测和故障量预测两个部分. 首先, 需要输入训练数据; 然后, 这些数据经过预处理形成训练模型, 训练模型就能对故障量进行预测; 最后就可输出预测结果.

2.2 数据预处理

由于原始数据存在缺失、错漏等问题, 因此需要先对原始数据进行预处理[9], 以得到可供算法输入的干净数据. 进行的数据预处理主要包括数据填充、特征规范化、特征映射等步骤.

(1) 数据填充

在实验的原始数据中不可避免地存在着一些异常数据和缺失数据, 这些数据若没有处理会影响到实验结果的准确性, 因此需要对这些数据进行处理. 常用的数据填充方法有均值填充、随机填充、线性回归填充以及 EM 填充等[10].

(2) 特征规范化

原始数据不同特征的值域可能存在较大差异. 例如: 在数据中, 有的特征的值域可能达到 1010 数量级, 而有的可能只有个位数. 如果直接在原始数据上分析, 数值大的特征将湮没数值小的特征, 使数值小的特征无法得到有效利用. 因此, 需要对原始数据做规范化处理. 常见的规范化方法有最小-最大规范化、z-score 规范化及小数定标规范化等[11].

(3) 特征映射

由于在原始数据中有的特征是以文字的形式来描述的, 因此需要转换为类别型数据才能作为算法输入. 而有的特征是数值型的数据, 在用于分类时也要转换为类别型数据才能作为算法输入. 因此我们可以定义一些转换规则来将特征进行映射. 常用的特征映射方法有自组织特征映射[12]、拉普拉斯特征映射[13]、等距特征映射[14]以及多域特征映射[15]等.

3 实验结果及分析

3.1 实验数据处理

该研究所用到的原始数据为上海各区 2012 年 1 月至 2015 年 5 月 21 日的天气情况及负荷数据等. 表 1 是区域负荷测点数据表所包含的字段, 表 2 是七天天气预报数据表所包含的字段.

数据预处理后就可以预测非故障量特征的值, 得到非故障量特征的值后就可进行故障量的预测. 数据进行的预处理操作如下:

(1) 数据填充

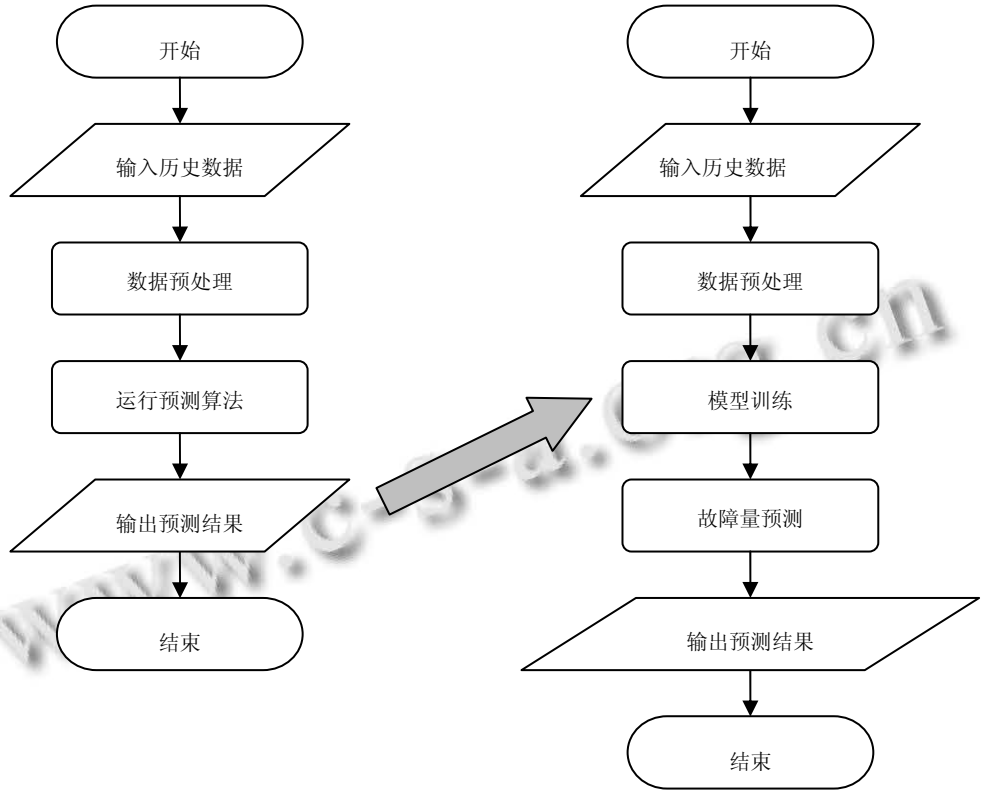
在本研究中, 对于原始数据中的缺失值, 将其填充为 0; 对于异常值, 由于包含异常值的记录很少, 直接删除包含异常值的记录.

(2) 特征规范化

在本研究中, 采用区间规范化的方法, 将所有特征值映射到[0,1]区间. 如果某个特征的取值全为 0, 不

对该特征规范化, 保持原始 0 值.

(3) 特征映射



(a) 非故障量特征预测　　(b)故障量预测

图 3　数据处理总体流程

本研究中所采用的转换规则如下:

① 天气情况的转换规则

在研究的原始数据中, 天气分为 95 种类型. 不仅类别过多, 且不同类别之间存在大量交叉, 如“阴~小雨”、“中雨~小雨”、“小到中雨~阵雨”等都存在重复的天气情况. 因此, 在保证合理性的基础上, 可将天气情况进行如下转换:

首先, 将天气分为雪、大雨、中雨、小雨、阴、多云和晴和其它等 8 类, 分别用数值 0~7 代表; 其次, 将 95 种天气归入符合的最坏的天气类型. 例如: “阴~小雨”归入小雨类, “中雨~小雨”归入大雨类, 等等. 这样, 划分后的天气情况可以突出坏天气对电网故障的影响.

表 1　区域负荷测点数据表

字段名	字段注释	字段类型
YMD	采集日期	NUMBER(10)
HMS	采集时间点	NUMBER(10)
POINT_ID	测点 ID	NUMBER(10)
FLAG	标示	NUMBER(4)
VALUE	测点值	NUMBER(38,6)

表 2　七天天气预报数据表

字段名	字段注释	字段类型
ID		NUMBER(10)
PUBLISH_DATE_TIME	发布日期	DATE
FORECAST_DATE_TIME	预测日期	DATE
BEGIN_WEATHER	开始天气情况	VARCHAR2(64)
END_WEATHER	结束天气情况	VARCHAR2(64)
HIGH_TEMP	最高温度	NUMBER(4)
LOW_TEMP	最低温度	NUMBER(4)
WIND_DIRECTION	风向	VARCHAR2(32)
WIND_VELOCITY	风力等级	VARCHAR2(32)

② 风力转换规则

风力转换规则如表 3 所示. 转换规则与天气情况

的转换规则类似, 将风力归纳为 1-2 级、3-4 级、5-6 级、7-8 级和其它等 5 个等级, 分别用数值 0~4 代表. 将原始风力情况归入相匹配的最大风力等级.

表 3 风力转换规则

序号	风力	数据转换规则
1	3-4 级	1
2	5-6 级	5
3	微风~3-4 级	1
4	4-5 级~3-4 级	2
5	微风	1
6	5-6 级~4-5 级	4
7	3-4 级~微风	1
8	3-4 级	1
9	4-5 级~3-4 级	2
10	4-5 级	3
11	3-4 级~4-5 级	2
12	4-5 级~5-6 级	4
13	微风~4-5 级	3

③ 故障量转换规则

由于实验数据中的故障量是数值型数据, 在用于分类时要转换为类别型数据才能作为算法输入, 因此采用故障量区间划分映射的方法将数值型数据转换为类别型数据.

首先, 采用 EM 算法对故障量进行区间的划分, EM 算法可以自动学习将故障量聚类成 K 个类簇. 故障量被聚为的类就是故障量被划分的区间. 然后, 将故障量原始值映射至各个类中. FAULT_COUNT_2_1、FAULT_COUNT_2_2 、 FAULT_COUNT_2_3 、FAULT_COUNT_2_4 指的是电压等级为 1、2、3、4 的故障量, FAULT_COUNT_2 是电网故障的总故障量, FAULT_COUNT_1 是非电网故障的总故障量.

FAULT_COUNT_2_1, FAULT_COUNT_2_2, FAULT_COUNT_2_3 和 FAULT_COUNT_2_4 的故障量区间划分和映射规则如表 4 所示; FAULT_COUNT_1 的故障量区间划分和映射规则如表 5 所示; FAULT_COUNT_2 的故障量区间划分和映射规则如表 6 所示.

表 4 故障量转换规则 1

区间	数据转换规则
[0,0]	1
[1,5]	2
[6,15]	3
[16,21]	4
[22,78]	5

表 5 故障量转换规则 2

区间	数据转换规则
[0,103]	1
[103,104]	2
[104,339]	3
[339.340]	4
[340,731]	5
[731,732]	6
[732,1731]	7
[1731,1732]	8
[1732,2748]	9
[2748,2749]	10
[2749,4713]	11
[4713,4714]	12
[4714,10128]	13

表 6 故障量转换规则 3

区间	数据转换规则
[0,0]	1
[0,1]	2
[1,5]	3
[5,6]	4
[6,17]	5
[17,18]	6
[18,92]	7

(4) 数据整合与重组

对原始数据进行整合, 生成新的整合数据表, 实验数据主要包含字段如表 7.

表 7 实验数据表

YMD	年月日
REGION_ID	区域 ID
BEGIN_WEATHER	一天初始时的天气情况
END_WEATHER	一天结束时的天气情况
WIND_VELOCITY	风力
RAIN_PROBABILITY	降雨概率(百分数)
HIGH_TEMP	最高气温
LOW_TEMP	最低气温
MAX_VALUE	负荷最大值
MIN_VALUE	负荷最小值
AVG_VALUE	负荷平均值
FAULT_COUNT_2_1	电压等级 1 的故障量
FAULT_COUNT_2_2	电压等级 2 的故障量
FAULT_COUNT_2_3	电压等级 3 的故障量
FAULT_COUNT_2_4	电压等级 4 的故障量
FAULT_COUNT_1	非电网故障的总故障量
FAULT_COUNT_2	电网故障的总故障量

3.2 实验结果分析

实验运用基于随机森林算法和基于 ELM 算法两种方法来进行故障量预测, 两种方法所用到的数据相同, 均是上海各区 2013 年 1 月至 2014 年 12 月 01 日的用电数据, 其中训练集中每个区域分别 300 条, 剩下的为测试集.

实验预测结束后需要对预测结果进行验证评估, 用到的是查准率和查全率两个指标[16]. 它们的计算公式如下:

$$precision = \frac{N_p}{N_t}$$

$$recall = \frac{N_p}{N_r}$$

其中, *precision* 指的是查准率, *recall* 指的是查全率, N_p, N_t, N_r 分别表示预测正确样本数、预测样本数及真实样本数.

表 2、表 3 是该实验结果的查准率和查全率, 表 1 是基于随机森林算法故障量预测结果, 表 2 则是基于 ELM 算法的故障量预测结果. 表中的 FAULT_COUNT_2_1 、 FAULT_COUNT_2_2 、 FAULT_COUNT_2_3、FAULT_COUNT_2_4 指的是电压等级为 1、2、3、4 的故障量, FAULT_COUNT_2 是电网故障的总故障量, FAULT_COUNT_1 是非电网故障的总故障量. 而表中的类别 0、类别 1、类别 2、类别 3、类别 4、类别 5 是在数据预处理时, 根据故障量转换规则, 对故障量进行的区间上的划分.

表 8　基于随机森林的实验结果分析表

RF		类别 0	类别 1	类别 2	类别 3	类别 4	类别 5
FAULT_COUNT_1	查准率(%)	0.00	53.92	74.66	97.17	0.00	0.00
	查全率(%)	0.00	58.35	81.25	90.11	0.00	0.00
FAULT_COUNT_2	查准率(%)	0.00	24.05	27.95	71.89	0.00	0.00
	查全率(%)	0.00	5.38	35.08	73.76	0.00	0.00
FAULT_COUNT_2_1	查准率(%)	0.00	100.00	0.00	0.00	0.00	0.00
	查全率(%)	0.00	99.63	0.00	0.00	0.00	0.00
FAULT_COUNT_2_2	查准率(%)	0.00	99.68	0.00	0.00	0.00	0.00
	查全率(%)	0.00	97.37	0.00	0.00	0.00	0.00
FAULT_COUNT_2_3	查准率(%)	0.00	0.00	18.77	86.04	0.00	0.00
	查全率(%)	0.00	0.00	26.70	79.25	0.00	0.00
FAULT_COUNT_2_4	查准率(%)	0.00	9.09	28.98	85.97	0.00	0.00
	查全率(%)	0.00	2.00	27.35	89.98	0.00	0.00

表 9　基于 ELM 的实验结果分析表

ELM		类别 0	类别 1	类别 2	类别 3	类别 4	类别 5
FAULT_COUNT_1	查准率(%)	97.87	1.11	0.00	0.00	0.00	0.00
	查全率(%)	11.08	41.67	0.00	0.00	0.00	0.00
FAULT_COUNT_2	查准率(%)	42.16	56.75	5.85	0.00	0.00	0.00
	查全率(%)	51.35	37.21	13.70	0.00	0.00	0.00
FAULT_COUNT_2_1	查准率(%)	98.74	0.00	0.00	0.00	0.00	0.00
	查全率(%)	99.97	0.00	0.00	0.00	0.00	0.00
FAULT_COUNT_2_2	查准率(%)	99.65	2.60	0.00	0.00	0.00	0.00
	查全率(%)	82.16	6.21	0.00	0.00	0.00	0.00
FAULT_COUNT_2_3	查准率(%)	65.23	35.09	0.59	0.00	0.00	0.00
	查全率(%)	56.26	39.18	1.56	0.00	0.00	0.00
FAULT_COUNT_2_4	查准率(%)	68.11	39.73	0.00	0.00	0.00	0.00
	查全率(%)	65.99	15.86	0.00	0.00	0.00	0.00

从表 8 和表 9 中可以看出, 随机森林算法的准确性高于 ELM 算法的准确性. 这是因为, 随机森林是以树的形式表示的规则集, 通过控制树深和树数以得到比较准确反映数据规律的预测结果. 因此, 当数据类别数较多时, 随机森林算法仍具有较好的精度和稳定性; 而 ELM 算法将划分规则隐含在网络权重中. 虽然 ELM 算法改进了传统神经网络算法多次迭代导致效率低的缺点, 但该算法的初始权重和隐层节点参数默认随机选取, 使得构建网络的稳定性难以得到保证, 容易造成分类结果精度的较大波动. 也就是说, ELM 算法处理类别较多的数据集时稳定性不足. 因此, 当要进行多分类的故障量预测时, 选择随机森林算法来进行故障量预测较为合适.

4 结论

本文将随机森林算法运用于配网抢修故障量预测中, 提出了一种基于随机森林算法的故障量预测方法. 该方法先预测了未来一天的天气、负荷等特征的值, 然后根据这些特征的值, 来预测未来一天的故障量. 经过数据集实验表明, 该方法能有效预测故障量. 此外, 实验中还运用 ELM 算法来构建故障量预测模型, 实验结果表明基于随机森林算法的预测结果相比 ELM 算法的预测结果具有更高的准确性. 随机森林算法不仅能够处理有噪声的数据, 而且不会过度拟合, 有良好的泛化性, 应用前景广阔. 基于随机森林算法的配网抢修故障量预测方法, 对配网抢修管理具有一定的参考价值. 下一步, 我们将研究如何细化故障量的区间范围, 以进一步提高故障量预测的精确度.

参考文献

1 郑琳.配网抢修精益化.中国电力企业管理,2011,(10):34–39.

2 袁仲雄.基于模糊评判法的配网抢修模型.华东电力,2011,39(2):249–252.

3 金家红,沈志宏,金良峰.基于移动数据终端的配网抢修系统的设计与应用.浙江省电力学会.浙江省电力学会2012年年会优秀论文集.中国浙江:中国电力出版社,2012.

4 张智敏.配电网故障抢修效率提升策略探析.电子测试,2013,(11):47–49.

5 蔡金锭,鄢仁武.基于小波分析与随机森林算法的电力电子电路故障诊断.电力科学与技术学报,2011,2:54–60.

6 Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32.

7 Breiman L. Bagging predictors. Machine Learning, 1996, (2): 123–140.

8 罗知林,陈挺,蔡皖东.一个基于随机森林的微博转发预测算法.计算机科学,2014,41(4):62–64.

9 方洪鹰.数据挖掘中数据预处理的方法研究[硕士学位论文].重庆:西南大学,2009.

10 邓银燕.缺失数据的填充方法研究及实证分析[硕士学位论文].西安:西北大学,2010.

11 蔡维玲,陈东霞.数据规范化方法对 K 近邻分类器的影响,2010,(22):175–178.

12 张义忠,赵明生.基于自组织特征映射的网页分类研究.信息与控制,2003,32(2):108–112.

13 蒋全胜,贾民平,胡建中,等.基于拉普拉斯特征映射的故障模式识别方法.系统仿真学报,2008,20(20):5710–5713.

14 陈法法,汤宝平,苏祖强.基于等距映射与加权 KNN 的旋转机械故障诊断.仪器仪表学报,2013,34(1):215–220.

15 窦万峰,王保保.多域特征映射机理研究与应用.机械工程学报,1998,34(5):34–39.

16 王洵.查全率与查准率.情报科学,1981,(3):40–44.