

控制与决策

Control and Decision

CART决策树方法在煤电厂节能降耗中的应用

张俊玉, 胡家豪, 黄嵩

引用本文:

张俊玉, 胡家豪, 黄嵩. CART决策树方法在煤电厂节能降耗中的应用[J]. *控制与决策*, 2021, 36(5): 1232–1238.

在线阅读 View online: <https://doi.org/10.13195/j.kzyjc.2019.1272>

您可能感兴趣的其他文章

Articles you may be interested in

[基于地标特征和元学习方法推荐最适用优化算法](#)

Recommending best suitable metaheuristic based on landmarking feature and meta-learning approach

控制与决策. 2021, 36(5): 1223–1231 <https://doi.org/10.13195/j.kzyjc.2019.0993>

[基于互信息操作变量曲线参数化的间歇过程批内修正优化](#)

Intra-batch correction optimization of batch process with manipulated variable trajectory parameterization based on mutual information

控制与决策. 2021, 36(1): 234–240 <https://doi.org/10.13195/j.kzyjc.2019.0825>

[社会网络环境下利益抗争的群策略研究](#)

Group strategy in struggling for interests under social network

控制与决策. 2021, 36(1): 206–215 <https://doi.org/10.13195/j.kzyjc.2019.0497>

[基于行为流图的可信交互检测方法](#)

Trustworthy interaction detection method based on user behavior flow diagram

控制与决策. 2020, 35(11): 2715–2722 <https://doi.org/10.13195/j.kzyjc.2018.1618>

[基于无标签、不均衡、初值不确定数据的设备健康评估方法](#)

Equipment health risk assessment based on unlabeled, unbalanced data under uncertain initial condition

控制与决策. 2020, 35(11): 2687–2695 <https://doi.org/10.13195/j.kzyjc.2018.1493>

CART 决策树方法在煤电厂节能降耗中的应用

张俊玉^{1†}, 胡家豪¹, 黄嵩²

(1. 中山大学 数学学院, 广州 510275; 2. 北京大学 软件与微电子学院, 北京 100871)

摘要: 煤电是我国电力的主要来源, 燃煤发电煤耗高, 所以降低发电厂的煤耗, 是非常重要且亟待解决的问题. 利用热力设备的大量监控数据, 提出利用 CART 决策树方法以及 CART 决策树与关联规则相结合的改进方法, 寻求不同工况下能降低单位煤耗的关键可控变量及其参数调节范围. 相比传统的用于寻找降耗调节方式的关联规则算法, CART 决策树方法无论在计算效率还是优化效果方面都具有一定的优越性, 适合在煤电厂实际操作中推广使用. 采用山东某热电厂锅炉的历史运行数据作实证研究, 发现用 CART 决策树方法给出的可控变量和它们的参数调节规则进行调控, 与现行的方法相比, 在各种工况下的煤耗节能百分比均超过 10%. 利用 CART 方法得到的调节规则下的煤耗比传统的关联规则方法也减少了 2.3%. 这个实例验证了 CART 决策树模型在煤电降耗的实际应用过程中的有效性. 最后, 为了使关键可控变量的调节范围更精确, 提出采用 CART 与关联规则相结合的改进方法.

关键词: 运行优化; 随机森林; 分类和回归树; 聚类分析; 降低煤耗; 特征选择

中图分类号: TM621

文献标志码: A

DOI: 10.13195/j.kzyjc.2019.1272

开放科学(资源服务)标识码(OSID):



引用格式: 张俊玉, 胡家豪, 黄嵩. CART 决策树方法在煤电厂节能降耗中的应用[J]. 控制与决策, 2021, 36(5): 1232-1238.

Application of CART decision tree model in reducing coal consumption in coal power plant

ZHANG Jun-yu^{1†}, HU Jia-hao¹, HUANG Song²

(1. School of Mathematics, Sun Yat-Sen University, Guangzhou 510275, China; 2. School of Software & Microelectronics, Peking University, Beijing 100871, China)

Abstract: The main source of electric power in China is coal-fired power, which consumes a large amount of coal. Therefore, reducing the coal consumption of power plants in the process of power generation is very important, which is urgent to be solved. In order to solve this problem, using a large number of data collected from thermal equipment, this paper proposes a CART method and a improved method combining the CART with association rules to find out the key controllable variables and their ranges that can reduce unit coal consumption under different working conditions. Compared with the traditional association rule algorithm used for energy saving, the CART method has higher calculation efficiency and achieves better optimizing effect. In addition, we use the CART to analysis the historical operation data of a thermal power plant in Shandong Province. It is found that under all cases of different motor power, the key controllable variables and their ranges given by the CART method can reduce the coal consumption by more than 10% compared with the current method. Compared with the traditional association rules, the regulation rules obtained using the CART method reduce the coal consumption by 2.3%. The analysis for the practical application of coal power consumption reduction rerifies the effectiveness of the CART model. Finally, in order to make the adjustment range of key controllable variables more accurate, an improved method combining And association rules is proposed.

Keywords: operation optimization; random forest; CART; cluster analysis; coal consumption reduction; feature selection

0 引言

煤电是我国电力供应中的基础性电源, 在我国发电总量中的占比超过 60%. 煤耗是反映煤电厂能耗

水平的核心指标之一, 它是指火力发电厂生产单位电能所耗用的燃料. 通俗地讲, 电厂煤耗就是燃煤电厂每生产单位千瓦时的电(俗称一度电)消耗的煤的

收稿日期: 2019-09-08; 修回日期: 2019-12-03.

基金项目: 国家自然科学基金项目(61673019).

责任编委: 阳春华.

[†]通讯作者. E-mail: mcszhjy@mail.sysu.edu.cn.

数量,单位是 $\text{g}/(\text{kW}\cdot\text{h})$ 。这里的煤是指标准煤,即发热量为 $7000\text{ kcal}/\text{kg}$ 的煤。国家能源局发布的《煤电节能减排升级改造行动计划(2014~2020年)》明确提出,2020年要实现现役燃煤发电机组改造后平均供电煤耗低于 $310\text{ g}/(\text{kW}\cdot\text{h})$,这一计划给电企节能降耗提出了明确要求。随着小机组的退役,30万kW和60万kW级机组的改造以及66万kW和100万kW新机组的投运,全国的平均煤耗逐年下降。但是当前的燃煤电厂平均供电煤耗却依然高达 $320\text{ g}/(\text{kW}\cdot\text{h})$ 的标准煤,说明热电厂的节能降耗仍然任重道远。

煤电厂的自动调节运行机制如下:用户处功率提升,则发电机转速下降,频率降低。这时电厂的反馈系统开始启动,汽轮机通过调节抽气等措施使动力增大,从而使转速提升至汽轮机额定转速,一般是3000转。在调节之后,汽机处由于提高了进气量,导致锅炉过热,蒸汽的压力降低,这时锅炉会通过调节给水量、给粉量、送风量、引风量、磨煤机转速等一系列参数使蒸汽压力回到正常值。这些调节过程的速度很快,电网电机和汽机调节也几乎是自动的。这种自动调节系统被广泛用于传统发电中。然而,采用自动调节也存在着一些局限。如果锅炉负荷变化很快,则自动调节系统会通过点油枪来满足蒸汽压力。但是,如果负荷提高过快而电厂发电量没有相应调整,则电网的频率就会降低,这不仅不能供给用户符合质量要求的电能,甚至会破坏整个生产链。因此,对于过快的电网负荷变化,采用自动调节的办法并不合适,人工调节的方式则显得尤为重要。人工调节参数主要靠经验,没有科学的指导方法,电网有时甚至采用突然改变负荷的方式训练电厂的应急能力。所以,针对不同的电网负荷,给出有科学依据的降低煤耗的参数调节方式,对于电厂的实际操作和运行具有重要意义。

机组发电的自动调节系统,本质上是一个PID控制器。但由于方程中的许多状态是不可观测的,从PID方程导出的最优化参数的可靠性存在质疑,因而难以建立控制系统的数理模型进行优化。同时,大型燃煤机组具有热力规模大、设备庞大、流量大、质量大等特点,在能量传递、转换和耗散等方面具有明显的非线性特征。比如,发电能耗与外部环境、资源和负荷需求之间就存在高度耦合和非线性关系^[1]。近年来,智能优化算法由于仅依赖系统的输入输出信息,在处理非线性系统问题上表现出显著优势^[2-3]。另一方面,在现代化热电厂运行中,电厂开始逐渐重视存储到数据仓库中的热力设备的监控数据。在这些大量数据背后,往往蕴涵着人工分析无法发现的规律,

因此,可以借助数据挖掘系统建模方法的优势,利用数据科学中的机器学习技术来研究电厂运行数据背后的规律性,提取出有效的信息,以解决电厂热力设备运行控制和优化问题^[4-6]。

国内外已经有很多学者使用数据挖掘的方法来提高燃烧效率。文献[7]等利用神经网络算法对锅炉的性能指标和一些非分析约束条件进行预测,从而加快了优化锅炉燃烧寻找最佳运行点的尝试过程。文献[8]开发了一个以排放物为研究目标的神经网络模型来优化锅炉的运行,从而减少了氮氧化物的排放,提高了锅炉的性能。但是,采用神经网络算法的一个劣势在于,它在优化控制设定点时,忽略了燃烧过程的时移性质。神经网络虽然是一种有效的建模技术,然而,这种方法训练速度慢,结果的可解释性差,并不适合实时优化的燃烧过程^[9]。所以,国外也有学者提出使用动态神经网络与静态神经网络模型相结合的方法^[10]研究电力工业领域的问题。

文献[11]应用关联规则算法挖掘电厂参数之间的关系,通过引入规则模拟专业系统,根据数据的层层筛选,获取与低煤耗有关的可调节变量。这种方法后来成为了国内研究数据驱动的燃烧效率优化的主要方法。此后,聚类算法、时间线性回归、决策树算法等也逐渐被用于挖掘火电厂参数之间的关系。近年来,有学者使用改进后的关联规则挖掘算法(如量化关联规则、模糊关联规则等)进行研究^[12-14],其基本思想是:首先将连续的数值变量划分成多个区间,转化成离散型变量;然后利用关联规则挖掘算法从历史运行数据中寻找锅炉效率的关联规则,从而得到参数的控制范围。但是,关联规则方法也存在一些不足:1)锅炉的运行参数基本都是连续的,而改进后的关联规则需要先将数据离散化。在变量离散化的过程中,绝大部分论文都使用均匀划分的方法,分成几类由人为设定、带有较强的主观性、没有客观的划分标准。2)在生成关联规则的过程中,有一些煤耗较低但是出现次数较少的参数组合,由于关联规则规定了最小支持度而未能进入频繁项目集,因此在最终生成的关联规则中可能失去了这部分效果极佳的参数组合。

用数据挖掘探索降低煤耗的参数调节方式,首先需要根据煤耗情况,分成高煤耗、中煤耗和低煤耗3类;然后再从低煤耗类别中确定可控变量的参数范围。虽然机器学习中有许多模型可以用来分类,但是大部分分类效果好的模型可解释性差,比如神经网络、支持向量机、随机森林等,尤其是近年来很热的深度学习模型。热力厂降低煤耗不光是个分类问题,更

是个控制优化问题,要求模型必须有可解释性. 这样可选的模型就少了.

决策树算法^[15]易于理解、可解释性强,可以对实际操作过程中大量复杂的可控参数进行特征选择,而且具有训练速度快、运行时间短的优点. 已有学者将决策树算法用于优化锅炉效率,如文献[16]. 但是该文献中使用的是C4.5算法,而C4.5算法也需要把连续属性转换为离散属性再进行处理,与之相比,CART(classification and regression trees)算法可以直接处理连续变量,因此,本文采用CART决策树算法进行研究,并以单位煤耗作为决策目标来寻找降低煤耗的直接控制因素.

在实际应用中决策树也有不稳定和容易过拟合两个主要缺点. 用随机森林的方法可以减小不稳定性. 随机森林算法通常能够收敛到更低的泛化误差,且具有简单、易实现、计算开销小等诸多优点. 但是由于随机森林方法使用了套袋(bagging: bootstrap aggregating)的思想,采用多个决策树作为机器学习器来投票的方法,随机森林没有办法说明变量是如何具体影响结果的,可解释性不强. 随机森林虽然不能解释变量如何影响结果,但可以给出变量的重要性排序. 于是,本文先用基于随机森林模型的递归特征消除法进行特征选择^[17],选择出重要的可控变量,避免全部特征放进模型,减少模型不稳定性 and 运行时间. 随机森林有两种常见的特征重要性评分,一种是平均不纯度减少(MDI),另一种是平均精确率减少(MDA). 文献[18]中指出,随机森林基于MDA的特征重要性排名对数据集的较小扰动是不稳定的,而基于MDI的特征重要性评分提供了更鲁棒的结果.

随机森林是由多个决策树构成的,决策树中的每一个节点都与某个特征有关. 对于分类问题,在训练决策树时,可以计算出每个节点上的特征减少了多少树的基尼不纯度. 对于某一随机森林而言,可以算出每个特征减少的基尼不纯度的平均值,并把这个平均值作为特征选择的评分值. 然而,使用基于平均不

纯度的重要性评分的方法,在样本数据不平衡时存在偏向问题,对容易分类至特征数量更多的类别的特征更有利,因而不能准确地反映特征的重要性^[18]. 所以,在使用特征重要性评分前,需要对不平衡的数据进行预处理.

有几种途径可以避免决策树学习中的过度拟合,它们可以被分为两类:一类是及早停止增长树法(早停法 early-stopping),在算法完美分类训练数据之前停止增长树;另一类是后修剪法(剪枝法 post-prune),即允许树过度拟合,然后对这个树后修剪. 本文选择限制叶子节点最小样本数的早停法,当一个叶子节点的样本数过小,则提前结束分裂,以此来克服过拟合的缺点. 这也是本文的创新点之一.

本文的主要实验步骤如下:第1步,按照主汽压力变化率小于5%的标准,筛选出锅炉运行处于稳态时的数据;第2步,按照电机功率划分工况;第3步,在每个工况下按照单位煤耗进行聚类,分成低煤耗、中煤耗和高煤耗3类;第4步,按照一定比例将上一步的数据分成训练集和测试集;第5步,判别是否存在不平衡问题,如果存在则使用SMOTE(synthetic minority over-sampling technique,即“人工少数类过采样法”,非直接对少数类进行重采样,而是设计算法来人工合成一些新的少数样本)过采样算法^[19]平衡训练集;第6步,在训练集中构建随机森林模型获得变量的重要性排序,据此剔除不重要的控制变量;第7步,构造决策树模型并进行参数优化;第8步,获得控制变量的参数范围;第9步,用测试集样本来验证该方法的有效性.

1 实验过程

1.1 数据来源

本文数据来源于山东某热电厂2017年4月至6月每30s实时锅炉运行的历史监控数据,包含110472个数据,存在且只有1个缺失值,采用删失的方式处理. 涉及变量包括时间序号、发电机功率、发电煤耗、26个可控变量和32个不可控变量,详见表1和表2.

表1 热电厂锅炉运行可控变量

序号	变量	序号	变量	序号	变量	序号	变量
1	给煤流量	8	引风机开度	15	三级抽气压力	22	凝结水温
2	过热器一级减温进出口温度差	9	一次风机电流	16	四级抽气压力	23	凝结水流量
3	过热器二级减温进出口温度差	10	一次风机开度	17	五级抽气压力	24	循环水泵电流
4	给水流量	11	二次风机电流	18	六级抽气压力	25	真空泵电流
5	一次风量	12	二次风机开度	19	七级抽气压力	26	冷却水泵电流
6	总风量	13	一级抽气压力	20	给水泵流量		
7	引风机电流	14	二级抽气压力	21	除氧器水温		

表2 热电厂锅炉运行不可控变量

序号	变量	序号	变量	序号	变量	序号	变量
1	主汽温度变化率	9	汽包压力	17	过热器二级减温口出口温度	25	空预器二次风温
2	主汽压力变化率	10	主汽压力	18	再热器出口温度	26	一级抽气温度
3	主汽过热度	11	再热器入口压力	19	给水管管温度	27	二级抽气温度
4	再热蒸汽过热度	12	再热器出口压力	20	气泡水位	28	三级抽气温度
5	中层床温	13	主蒸汽温度	21	炉膛负力	29	四级抽气温度
6	下层床温	14	过热器一级减温口入口温度	22	预热器入口处烟气氧量	30	五级抽气温度
7	床压	15	过热器一级减温口出口温度	23	排烟温度	31	六级抽气温度
8	床层压差	16	过热器二级减温口入口温度	24	空预器一次风温	32	七级抽气温度

1.2 数据预处理

1) 数据筛选. 锅炉刚开始加热的过程是一个非稳态过程,在该过程中产生的数据样本可能会对结果产生不良影响,因此需要提取稳态工况下的数据进行分析. 根据文献[12],当锅炉内的主汽压力和主汽温度变化不大时,可以认为此时锅炉的运行工况是稳定的. 由于上述二者之间存在显著的正相关关系,本文界定主汽压力变化率小于5%的数据为锅炉运行稳态时的数据,并以该数据作为本文的研究对象.

2) 变量选择. 本文研究目的是找到能使煤耗降低的变量取值范围. 本文研究问题不但是预测问题,而且是控制问题. 原始数据中的32个不可控变量只是反映当前锅炉运行情况的监控数据,它们的取值是不能控制的,这些变量对控制问题的研究作用不大. 因此,本文将32个不可控变量剔除,仅对发电机功率、发电煤耗和26个可控变量进行数据分析.

3) 工况分类. 按照发动机功率(负荷)对工况进行分类. 数据样本中发动机功率取值范围是10MW~140MW之间. 本文将发电机功率从10MW开始,每10MW为一类,本文主要以发电机功率在100~110MW的工况为例进行分析,这个工况的数据约占到稳态数据的31%,为所有稳态数据中占比最高的工况.

1.3 煤耗的划分

根据发电机功率在100~110MW稳定工况下的32665个数据,可以得到电机功率在100~110MW的工况下发电煤耗的平均值为288.72g/(kW·h)时,中位数为274.13g/(kW·h). 本文的研究目的是寻找可以降低煤耗的规则,对发电煤耗没有细致的分类要求,因此采用mini batch K-means聚类分析算法[20],设置簇为3,将煤耗分为3类,从而将发电煤耗离散化.

Mini Batch K-means聚类分析的结果显示,将发电煤耗分为3类后,其中第1类的聚类中心为260.78g/(kW·h),此类区间下界为125.27g/(kW·h),上界为282.68g/(kW·h),此类区间整体低于全部发电

煤耗数据的平均水平,平均煤耗为260.46g/(kW·h),比整体平均值288.72g/(kW·h)还低9.8%,且符合《煤电节能减排升级与改造行动计划(2014~2020年)》提出的目标平均煤耗,因此将此类归为“低煤耗”,记为“L”类. 同理,将第2类和第3类分别归为“中煤耗”,“高煤耗”类,分别记为“M”和“H”类,如表3所示.

表3 100~110MW发电煤耗聚类情况 g/(kW·h)

划分类别	聚类中心	聚类区间
低煤耗类(“L”类)	260.78	[125.27, 282.68]
中煤耗类(“M”类)	304.59	[282.69, 341.09]
高煤耗类(“H”类)	377.62	[341.11, 643.85]

1.4 数据不平衡的处理

在该32665数据中,首先按照4:1的比例划分训练集和测试集,其中26132条数据作为训练数据,6533条数据作为测试数据. 训练集中包含16230条“L”类数据,5504条“M”类数据,4398条“H”类数据. 由于样本数目不平衡,样本数较少的一类分类效果可能较差,且决策树对于样本不平衡的分类问题处理不佳. 所以,本文使用SMOTE算法[19]进行过采样,填充“M”和“H”的样本,得到一个新的训练集,包括3类样本数据各16230条.

1.5 特征选择

对这26个可控变量与发电煤耗建立随机森林模型,为了使得到的模型特征重要性的结果更加稳定,设定随机森林包含的决策树个数为100个,得到基于平均不纯度的各个特征重要性排序. 本文设置0.02的阈值以筛选变量,最终剩下21个重要特征.

1.6 决策树的建立

用筛选后的数据建立决策树模型,按照4:1划分训练集和测试集,其中48760条数据作为训练数据,6533条数据作为测试数据.

决策树容易造成过拟合,一般通过剪枝来提高泛化能力. 有两种常用的剪枝方法:先剪枝(prepruning)和后剪枝(postpruning). 先剪枝是通过提前停止树的构建而对树“剪枝”,包括设置决策树的最大深度,设

置叶子节点的最小样本数,设置每次扩张对系统性能增益的最小值等.后剪枝是首先构造完整的决策树,允许树过度拟合训练数据,然后对那些置信度不够的结子树用叶子结点来代替.先剪枝方法不需要生成整个决策树,相对简单,效率较高,适合于解决大规模问题.由于本文的实验数据较大,所以采用先剪枝的方法.

在实验中发现,设置决策树的最大深度和设置叶子节点的最小样本数结果差别不大.由于叶子节点的最小样本数的参数选择范围比较大,以下只考虑设置决策树的最大深度,使得数据被划分为“L”类的准确率尽可能提高.使用Python中的Sklearn包学习建立决策树^[21].经过尝试,当决策树最大的深度设为13时,“L”类的准确率达到最高,因此本文将决策树最大深度设为13(决策树中参数设置max depth = 13).

使用测试集对决策树的预测结果进行验证,得到该决策树模型的准确率约为90.71%,其中“L”类的查准率约为93%,说明拟合较为良好.决策树部分分支如图1所示.

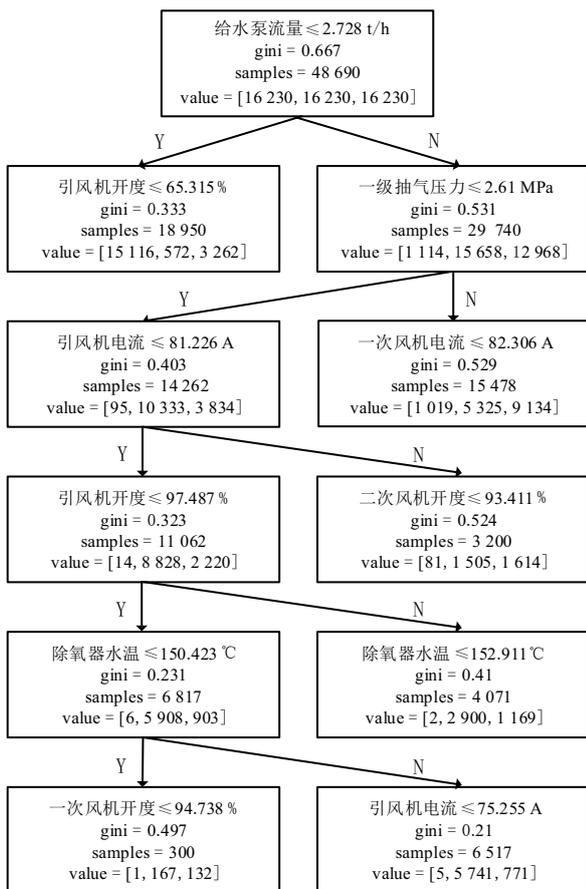


图1 决策树分支

根据“L”类分支的数据,可以获得在发电机功率为100~110MW的工况下,能够降低煤耗的控制方法,根节点到最终的叶节点所经历的路径上的每个叶节点上的规则即本文所需的结果.在其他工况,也可

以用类似的方法得到可以降低煤耗的调节规则.

2 实验结果分析

2.1 CART决策树算法的实验结果

由该决策树模型可找到一个分支,在该叶节点上样本容量为6517,分别包含5741个“L”类样本,5个“H”类样本,771个“M”类样本,基尼指数很小,为0.21,代表此时分类效果已经较佳.由该叶子节点可以看出,16230个低煤耗样本中有5741个样本被分到该节点上,说明分到该节点的规则具有很好的普遍性且不包含冗余的规则.截取到这个节点不再往下分析,该节点以上的规则即本文所寻找的可以降低发电煤耗的主要方式.由此可以根据决策树树形图的路径生成规则,将部分规则列举如表4.

表4 采用决策树算法得到的调节规则

可控变量	给水泵流量	一级抽气压力	引风机电流	除氧器水温
调节方式	>2.728 t/h	<2.61 MPa	<81.266 A	> 150.423 °C

当发电机功率处于100~110MW的工况下时,首先考虑增加给水泵流量.给水泵是发电厂最主要的辅助设备之一,其稳定性程度直接影响机组运行,当锅炉给水泵的流量小于要求的流量时,流过给水泵的给水会局部汽化,产生汽蚀而损坏给水泵.控制给水泵流量使其大于2.728 t/h,既可以保证机组的安全运行又可以降低煤耗.其次考虑调节抽汽系统中的一级抽气压力,使其低于2.61 MPa.其他可以采纳的建议如引风机电流小于81.266 A,除氧器水温大于150.423 °C.

按照以上规则在测试样本里共筛选出2286个样本,其平均煤耗为259.18 g/(kW·h),远低于全部测试集样本的平均煤耗288.72 g/(kW·h)时,节能10.2%.

在其他工况的时候,也可以用类似的方法得到要调整的特征及其范围,将其总结如表5所示.由于发电机功率在120~140MW之间的样本比较少,所以划分成一类.由表5还可以看出,发电功率在90~100MW时,煤耗减低较少,这是因为在这个工况下,测试集的整体平均煤耗本来已经很低,下降空间较少.但是,无论在何种工况下,发电煤耗都减少超过10%,证实了本文研究方法的有效性.

2.2 改进:离散CART决策树算法

在上述CART决策树模型中,也存在一些不足.其最主要的问题是CART的基础是二叉树,简单地将数值一分为二,得到的规则通常是可控变量的一个无限区间.譬如上述CART决策树的第1条规则,给水泵流量大于2.728 t/h,但在实际操控中这个范围太大,

表5 筛选前后平均煤耗对比表

工况/MW	平均煤耗(筛选前)/(g/(kW·h))	平均煤耗(筛选后)/(g/(kW·h))	节约百分比/%	测试集样本个数
80~90	304.56	226.08	226.08	1732
90~100	260.10	232.67	10.54	3811
100~110	288.72	259.18	10.25	6533
110~120	327.72	281.18	14.20	5031
120~140	410.49	314.87	23.29	3477

对于控制问题而言,希望得到一个更精确的范围.

对采样后的训练样本数据使用与关联规则算法相同的方式,将各个变量聚类离散化(每个特征都聚成5类),再用离散化的变量建立CART决策树(离散CART决策树算法),选取部分结果如表6所示.将CART方法、离散CART方法与传统的关联方法得到

的结果进行对比,如表7所示.

表6 采用离散CART决策树算法得到的调节规则

可控变量	给水泵流量/(t/h)	一级抽气压力/MPa	二次风机开度/%
控制区间	[2.78, 2.95] 或[2.95, 3.32]	[2.40, 2.52] 或[2.52, 2.58]	[71.11, 83.17] 或[89.46, 98.95]

表7 关联规则方法、CART决策树方法、离散CART决策树方法结果对比

指标	关联规则	决策树	离散化决策树
调节变量	给水泵流量 循环水泵电流 除氧器水温 二级抽气压力 三级抽气压力	给水泵流量 一级抽气压力 引风机电流 除氧器水温	给水泵流量 一级抽气压力 二次风机开度
运行时间/s	77.01	1.73	3.30
平均煤耗(筛选后)/(g/(kW·h))	268.70	259.18	260.06
节约百分比/%	7.9	10.2	10.2

根据对比结果可以发现:1) CART决策树的训练速度相较于关联规则大大提高,节约了大量运行时间;2)将关联规则和决策树得到的规则对比,两者均存在给水泵流量,除氧器水温变量等参数限定,结论差别不大,而且它们都将给水泵流量列为首要调节对象,因此可以认为,给水泵流量是调节发电煤耗的关键因素;3)本文在测试集上分别使用关联规则中最大置信度的关联规则和取决策树中基尼指数较小的一个节点的方式筛选数据,得到的平均煤耗为268.70 g/(kW·h),对比两种决策树分别为259.42 g/(kW·h)和260.06 g/(kW·h),决策树规则所得到的煤耗更低,节约百分比比关联规则高约2.3%;CART方法的运行时间是1.73 s,离散CART决策树方法的运行时间是3.30 s,只多了1.57 s,在实际中是可以接受的.况且他们的平均煤耗分别是259.18 g/(kW·h)和260.06 g/(kW·h),几乎没有差别.用1.73 s的代价换取更精确的参数范围在实际应用中是可取的.

3 结论

本文提出使用CART决策树方法和CART决策树与关联规则相结合的改进方法将煤电厂单位煤耗

作为目标变量建立模型,寻找直接影响煤耗的可控制变量,为煤电厂节能降耗的难题提供解决思路.在模型选择上,采用CART决策树算法,该算法不仅具有较强的可解释性,可以直接处理连续型数据,而且训练速度快,运行时间短,相比以往使用广泛的关联规则算法等具有显著的优越性.在细节处理上,本文使用了基于随机森林特征重要性的过滤式方法对特征进行筛选,减少决策树的不稳定性,防止得到影响并不大的变量调节规则,并以发电煤耗低的分类的查准率作为最终目标进行调参,最终发现使用CART决策树方法得到的调节规则可以使100~110 MW工况下的发电煤耗均降低10.2%,比传统的关联规则方法得到的调节规则还低2.3%,在其他工况下节能百分比也均超过10%,充分证实了CART决策树方法在煤电厂节能降耗中的有效性.

最后,本文将CART决策树方法改进,即先将可控变量聚类离散化,再使用CART决策树方法建立模型,最终提出了使用CART决策树与关联规则相结合的方法,该方法从计算效率和优化效果上均优于传统关联规则,在煤电厂节能降耗的实际运用中具有重要的指导意义.

参考文献(References)

- [1] Yang Y P, Wang N L, Zhang Z W, et al. Data mining-based modeling and application in the energy-saving analysis of large coal-fired power units[C]. International Conference on Machine Learning & Cybernetics. Qingdao: IEEE, 2010: 1095-1100.
- [2] 范永青, 王文庆, 江祥奎, 等. 一类不确定非线性离散系统的模糊自适应控制器设计[J]. 控制与决策, 2019, 34(3): 542-548.
(Fan Y Q, Wang W Q, Jiang X K, et al. Fuzzy adaptive control design for a class of uncertain nonlinear discrete-time systems[J]. Control and Decision, 2019, 34(3): 542-548.)
- [3] 樊兆峰, 马小平, 邵晓根. 非线性系统RBF神经网络多步预测控制[J]. 控制与决策, 2014, 29(7): 1274-1278.
(Fan Z F, Ma X P, Shao X G. RBF neural network multi-step predictive control for nonlinear systems[J]. Control and Decision, 2014, 29(7): 1274-1278.)
- [4] 卢勇, 徐向东, 陈明. 数据挖掘技术在热电厂过程控制与优化中的应用研究[J]. 电站系统工程, 2003, 19(2): 48-50.
(Lu Y, Xu X D, Chen M. Use of data mining techniques in process control and optimization of thermal power plant[J]. Power System Engineering, 2003, 19(2): 48-50.)
- [5] 孙秋野, 杨凌霄, 张化光. 智慧能源——人工智能技术在电力系统中的应用与展望[J]. 控制与决策, 2018, 33(5): 938-949.
(Sun Q Y, Yang L X, Zhang H G. Smart energy—Applications and prospects of artificial intelligence technology in power system[J]. Control and Decision, 2018, 33(5): 938-949.)
- [6] 胡家豪. 数据挖掘在煤电厂降低煤耗中的应用[D]. 广州: 中山大学数学学院, 2019.
(Hu J H. Application of data mining in reducing coal consumption in coal cower clant[D]. Guangzhou: School of Mathematics, Sun Yat-Sen University, 2019.)
- [7] Chu J Z, Shieh S S, Jang S S, et al. Constrained optimization of combustion in a simulated coal-fired boiler using artificial neural network model and information analysis*[J]. Fuel, 2003, 82(6): 693-703.
- [8] Booth R C, Roland W B. Neural network-based combustion optimization reduces NOx emissions while improving performance[C]. Dynamic Modeling Control Applications for Industry Workshop. Vancouver: IEEE, 1998: 1-6.
- [9] Kusiak A, Song Z. Combustion efficiency optimization and virtual testing: A data-mining approach[J]. IEEE Transactions on Industrial Informatics, 2006, 2(3): 176-184.
- [10] Selvi B T, Kalpana D, Thyagarajan T. Modeling and prediction of boiler drum in a thermal power plant[C]. Trends in Industrial Measurement and Automation (TIMA). Chennai: IEEE, 2017: 1-6.
- [11] Ogilvie T, Swidenbank E, Hogg B W. Use of data mining techniques in the performance monitoring and optimisation of a thermal power plant[C]. IEE Two-day Colloquium on Knowledge Discovery and Data Mining. London: IET, 1998: 7.
- [12] 冯春晖, 陈彦桥, 刘金琨. 数据挖掘技术在火电机组运行参数优化中的应用[C]. 第30届中国控制会议. 烟台, 2011: 2119-2123.
(Feng C H, Chen Y Q, Liu J K. Operating parameters optimization of thermal power unit based on data mining[C]. The 30th Chinese Control Conference. Yantai, 2011: 2119-2123.)
- [13] 李建强, 刘吉臻, 张栾英, 等. 基于数据挖掘的电站运行优化应用研究[J]. 中国电机工程学报, 2006, 26(20): 118-123.
(Li J Q, Liu J Z, Zhang L Y, et al. The research and application of fuzzy association rule mining in power plant operation optimization[J]. Proceedings of the CSEE, 2006, 26(20): 118-123.)
- [14] 王秋平, 陈志强, 魏浩. 基于数据挖掘的电站运行参数目标值优化[J]. 电力科学与工程, 2015, 31(7): 19-24.
(Wang Q P, Chen Z Q, Wei H. The summary of optimal operation parameters in power station based on the data mining[J]. Electric Power Science and Engineering, 2015, 31(7): 19-24.)
- [15] Rutkowski L, Jaworski M, Pietruczuk L, et al. The CART decision tree for mining data streams[J]. Information Sciences, 2014, 266: 1-15.
- [16] 王文欢, 潘秉超, 潘卫国. 基于决策树的燃煤锅炉运行优化规则的提取[J]. 上海电力学院学报, 2013, 29(2): 151-154.
(Wang W H, Pan B C, Pan W G. Research on rule extraction of coal-fired boiler optimizing operation based on decision tree[J]. Journal of Shanghai University of Electric Power, 2013, 29(2): 151-154.)
- [17] Dash M, Liu H. Feature selection for classification[J]. Intelligent Data Analysis, 1997, 1(1/2/3/4): 131-156.
- [18] Calle M L, Urrea V. Letter to the editor: Stability of random forest importance measures[J]. Briefings in Bioinformatics, 2011, 12(1): 86-89.
- [19] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
- [20] Sculley D. Web-scale k -means clustering[C]. Proceedings of the 19th International Conference on World Wide Web. Raleigh: ACM, 2010: 1177-1178.
- [21] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12(10): 2825-2830.

作者简介

张俊玉(1976—), 女, 副教授, 博士, 从事随机优化、离散事件动态系统等研究, E-mail: mcszhjy@mail.sysu.edu.cn;

胡家豪(1995—), 男, 硕士生, 从事数据挖掘、机器学习的研究, E-mail: 767101726@qq.com;

黄嵩(1977—), 男, 副教授, 博士, 从事互联网金融等研究, E-mail: huangsong@ss.pku.edu.cn.

(责任编辑: 孙艺红)