

Short-term Power Load Forecasting and Load Interval Prediction Method Based on Data Mining

Xiangxun Sun

Dongfang Electric Group International Cooperation Co., Ltd., Chengdu, Sichuan, 610017, China

Abstract

In order to make full use of the valuable historical data of power load data, this paper proposes a short-term power load prediction method based on data mining, which combines statistical strategies with machine learning to dynamically improve the prediction accuracy. Historical power load feature data set and mapping table form the basis of data mining. Based on statistical strategies, the historical similar day sequence is constructed, and the similar day power load sequence is mapped. As the training sample, it is input into the machine learning and training model, and the optimal prediction coefficient is obtained through the model training, and the short-term power load prediction is completed. The fluctuation range of power load prediction is also of great practical significance. This paper also proposes an innovative mode for research.

Keywords

power load; data mining; short-term load forecast; interval prediction; prediction method

基于数据挖掘的短期电力负荷及负荷区间的预测方法

孙湘珣

东方电气集团国际合作有限公司, 中国 · 四川 成都 610017

摘 要

为充分利用宝贵的电力负荷数据历史资料, 论文提出一个基于数据挖掘的短期电力负荷预测方法, 其中将统计策略与机器学习相结合以动态提升预测精度。历史电力负荷特征数据集和映射表构成了数据挖掘的基础, 基于统计策略构建历史相似日序列, 映射相似日电力负载序列, 作为训练样本输入机器学习和训练模型中, 经过模型训练获得最优预测系数, 完成短期电力负荷预测。电力负荷预测的上下波动范围也极具现实意义, 论文也提出一种创新性模式供研究。

关键词

电力负荷; 数据挖掘; 短期负荷预测; 区间预测; 预测方法

1 引言

电力负荷预测在现代电力系统中具有重大意义, 负荷预测可以帮助电力系统调度部门提前制定发电计划保障电力供应的稳定性, 合理分配资源, 避免供需不平衡。准确的负荷预测有助于优化电力系统的运行和调度, 降低从发电厂到电网的综合运营成本, 通过合理安排发电计划, 避免过剩或不足的发电, 可以减少燃料成本、运输成本等。随着“双碳”目标的推进, 构建以新能源为主体的新型电力系统是必然趋势。负荷预测对于新型电力系统的规划、运行和控制具有关键作用, 有助于实现清洁能源的高效利用。

电力预测是电力系统运行和规划中的关键环节之一, 已经在电力系统中得到广泛应用, 包括负荷预测、风电和太阳能发电预测、电价预测等。电力预测已经发展出种类繁多

的方法, 创新的模式和技术还在不断涌现, 以满足不断变化的需求和挑战。基于人工智能的预测模型: 利用深度学习、神经网络、卷积神经网络 (CNN) 和循环神经网络 (RNN) 等人工智能技术, 可以构建更准确、自适应的负荷预测模型。多源数据融合: 将不同来源的数据 (如天气数据、社交媒体数据、经济指标等) 与传统负荷数据融合, 以提高预测精度。实时预测和调度: 随着电力系统的智能化和数字化, 实时负荷预测和调度成为关键。新型模型和算法可以更快速地适应实时数据。可解释性和不确定性处理: 新型模型应该具备可解释性, 以使用户理解预测结果。电力负荷预测领域仍然具有许多创新的机会。

2 提出一个新的数据挖掘模式

电力运行是人类社会运行的一部分, 在时间上遵循一样的周期, 与自然的运行周期深度融合, 历史总是一再重复, 总体相似, 足以借鉴。

夏至和冬至是自然周期中的两个重要节气, 对人类社

【作者简介】孙湘珣 (1972-), 女, 中国湖南邵阳人, 本科, 工程师, 从事计算机技术应用、企业数字化转型研究。

会的运行有着深远影响。夏至：通常发生在每年的 6 月 21 日左右，夏至标志着盛夏的到来，天气炎热，植物生长茂盛，天文学上，夏至是北半球白昼最长的一天，太阳直射北回归线，阳光最为充沛。冬至：通常发生在每年的 12 月 21 日左右，冬至标志着冬季的开始，白昼最短，黑夜最长。天文学上，冬至是北半球白昼最短的一天，太阳直射南回归线。夏至和冬至作为自然周期的标志，也是人类社会生活的周期标志，当然电力负荷运行也与之高度重合。

步骤一：以冬至之日起始建立每一年周而复始的日期序列，提取历史记载中对应日期的温度、湿度形成向量元，构成第一个矩阵 A，同时将当日对应的负荷记录形成矩阵 B，得到两组数据的矩阵。

$$\mathbf{A} = \begin{bmatrix} D_{y(-1)1} & D_{y(-1)2} & \cdots & D_{y(-1)i} \\ D_{y(-2)1} & D_{y(-2)2} & \cdots & D_{y(-2)i} \\ \vdots & \vdots & \ddots & \vdots \\ D_{y(-k)1} & D_{y(-k)2} & \cdots & D_{y(-k)i} \end{bmatrix} \quad (1)$$

$$\mathbf{B} = \begin{bmatrix} V_{y(-1)1} & V_{y(-1)2} & \cdots & V_{y(-1)i} \\ V_{y(-2)1} & V_{y(-2)2} & \cdots & V_{y(-2)i} \\ \vdots & \vdots & \ddots & \vdots \\ V_{y(-k)1} & V_{y(-k)2} & \cdots & V_{y(-k)i} \end{bmatrix}$$

$D_{y(-k)i} = (i, t, h)$ 它是矩阵 A 中的一个向量元，表示过去第 k 年的以冬至日为起始日的第 i 日，当日的平均气温值 t，平均湿度值 h； $V_{y(-k)i}$ 是当日的实际电力负载数值

步骤二：设定今日之前一日的向量元： $D_{y(0)(i-1)} = ((i-1), T_{y(0)(i-1)}, H_{y(0)(i-1)})$ ，相应的负载记录 $V_{y(0)(i-1)}$ ；通过相似性计算获得今年之前 k 年中 A 矩阵中每年的最相似向量元 $(D_{y(-1)(i-1)best}, D_{y(-2)(i-1)best}, \cdots, D_{y(-k)(i-1)best})$ ，这一组向量元对应获得 B 矩阵中的一组样本，定义为历史相似量序列：

$$(V_{y(-1)(i-1)best}, V_{y(-2)(i-1)best}, \cdots, V_{y(-k)(i-1)best}) \quad (2)$$

也就是过去 k 年中历史记载的相似日负荷值的记录序列。

步骤三：针对历史相似量序列，采用指数移动平均 (EMA) 方法计算获得前一日的计算负荷值，同理依次获得前 n 日的完整计算负荷值序列：

$$(VC_{y(0)(i-1)}, VC_{y(0)(i-2)}, \cdots, VC_{y(0)(i-n)}) \quad (3)$$

步骤四：与此计算负荷值序列对应的当日实际负荷记录序列，如下 (4)，采用均方根误差 (RMSE) 方法计算这两个序列的误差值 RMSE，用以评价其相似度：

$$(V_{y(0)(i-1)}, V_{y(0)(i-2)}, \cdots, V_{y(0)(i-n)}) \quad (4)$$

步骤五：可以直接定义这组数列的 n 日为学习集，采用机器学习模式对学习集进行拟合运算，使得实际负荷记录序列与计算负荷值序列之间的均方根误差 (RMSE) 最小化，这样过去 n 日中每一日获得一个最优 α ，也即构成最优 α 序列，如下 (5)，作为学习结果，将 n 日的 α 算术平均值定义为当日最优 $\alpha: \alpha_{y(0)(i)best}$ ，用以对下一日的预测计算：

$$(\alpha_{y(0)(i-1)}, \alpha_{y(0)(i-2)}, \cdots, \alpha_{y(0)(i-n)}) \quad (5)$$

步骤六：通过天气预报，获得下一日的向量元参数：

$D_{y(0)(i+1)} = ((i+1), T_{y(0)(i+1)}, H_{y(0)(i+1)})$ ，重复上述步骤二至五，采用当日最优 α 在指数移动平均 (EMA) 方法中获得次日的，从而实现对次日的预测。

3 历史数据的相似性运算

步骤一中的 A、B 矩阵①，本质上是根据模糊聚类规则构建了电力负荷特性指标数据的模糊系数特征样本空间和映射表，构成负荷数据挖掘的基础，有效降低负荷随机性，保障挖掘精度。样本空间的边界取决于历史数据的累积，同时在实际运算中也可以根据精度的满意度进行合理切割，确保运算效率和精度的平衡。

今年第 i 日的向量为 $D_{y(0)(i)} = (i, T_{y(0)(i)}, H_{y(0)(i)})$ ，根据冬至日起始的周年循环推理，与其气温、湿度特征值最为相似的历史记录日应为过去每年的当日，即发生于：

$$(D_{y(-1)(i)}, D_{y(-2)(i)}, \cdots, D_{y(-k)(i)})$$

实际上每年的具体情况不同，因而最有可能的相似日是发生在日的前后数日之内，具体判定方法可以采用向量空间中的最短距离计算，即空间三维坐标的两点，怎么计算它们的距离？当已知一个点 $d_1 = (x_1, y_1, z_1)$ ，周围有若干个其他点 $d_i = (x_i, y_i, z_i)$ 时，我们可以计算这些点与已知点之间的距离，并找到最靠近的点。使用欧几里得距离公式计算它与已知点 D1 之间的距离：

$$d_i = \sqrt{(x_i - x_1)^2 + (y_i - y_1)^2 + (z_i - z_1)^2} \quad (6)$$

遍历所有点，找到具有最小数值的点，这个点就是距离已知点最近的点，因此我们可以用以下公式来表示选择距离最近的点：

$$d_i = \operatorname{argmin}_d \|d - d_1\| \quad (7)$$

根据坐标，获得 (1) B 矩阵中的数值 $V_{y(-k)i}$ 为最佳值，形成历史相似量序列：

$$(V_{y(-1)(i-1)best}, V_{y(-2)(i-1)best}, \cdots, V_{y(-k)(i-1)best}) \quad (8)$$

4 关于 EMA 运算

指数移动平均 (EMA) 是一种常用的平滑方法，用于对序列数据进行加权平均。它在深度学习中优点突出：

①平滑数据：EMA 会给近期的数据点赋予更大的权重，而对较早期的数据点赋予较小的权重。这有效地平滑时间序列数据，使其更连续和稳定。

②稳定模型参数更新：在深度学习中，EMA 通常用于平滑模型参数的更新。每次更新参数时，对模型参数进行 EMA 处理，减少每次更新的波动，使模型更加稳定。

③计算滑动平均梯度：EMA 还可以用于计算滑动平均

梯度,用于优化器的更新,进一步提高模型的性能和泛化能力。

以年度周期递进的电力负载向量天然构成时间序列,指数移动平均 (EMA) 随时间梯度由近及远平滑权重的特性完美反应样本序列的自然属性,地球宏观气候变化、各地区经济发展趋势等一系列长期影响因素被包容在当日最优 α ($\alpha_{y(0)(i)best}$) 的持续优化进程中。

EMA 的计算公式为:

$$ema_{(t)} = \alpha \cdot X_{(t)} + (1 - \alpha) \cdot ema_{(t-1)} \quad (9)$$

其中, t 为时间步; $X_{(t)}$ 为第 t 个时间点的原始数据; α 为平滑因子,通常取值在 0 到 1 之间,为当前样本的权重; $(1 - \alpha)$ 则为历史数据的权重; $ema_{(t-1)}$ 为上一个时间点的 EMA 值。通过 EMA,我们对历史数据进行加权平均,使得每个数据点的权重随着其距离当前时间点的远近而不断减小。这有助于更好地捕捉数据的趋势和变化。

5 关于均方根误差 (RMSE) 评价方法

均方根误差 (Root Mean Squared Error, RMSE) 是一种常用的评估预测模型精度的指标,直观性强: RMSE 的单位与原始数据的单位相同,误差度量 (如 MSE) 更容易解释和理解; 对大误差敏感: 由于 RMSE 计算的是误差的平方根,它对较大的误差更加敏感,对那些偏差较大的预测值给予更大的权重,有助于识别和改进模型中的大误差; RMSE 在许多领域都有广泛应用,包括气象预测、金融分析、工程设计等。它可以帮助评估模型的预测性能,并在不同模型之间进行比较。

最佳实用场景包括: 机器学习模型在时间序列预测中进行回归分析。

在机器学习中, RMSE 常用于评估回归模型的性能,帮助选择最佳的模型和调整超参数,特别是在需要对未来数据进行精确预测的情, RMSE 可以用来评估模型的训练效果,并进行模型的选择和优化。较小的 RMSE 值通常表示模型具有更好的泛化性能。

第一组样本 (3), 计算负荷值序列:

$$VC_{y(0)(i-1)}, VC_{y(0)(i-2)}, \dots, VC_{y(0)(i-n)}$$

第二组样本 (4), 实际负荷记录序列:

$$V_{y(0)(i-1)}, V_{y(0)(i-2)}, \dots, V_{y(0)(i-n)}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (VC_{y(0)(i)} - V_{y(0)(i)})^2} \quad (10)$$

其中, $VC_{y(0)(i)}$ 为第一组样本中的第 i 个值; $V_{y(0)(i)}$ 为第二组样本中的第 i 个值; n 为样本的数量。

这个公式表示的是两组样本对应值之间的平方差的平均值的平方根。RMSE 越小, 表示两组样本之间的误差越小。

6 关于机器学习的迭代最优化算法

机器学习中的迭代最优化回归算法是指通过反复调整

模型参数, 以最小化损失函数 (如均方误差) 的方法。这些算法通常利用梯度信息来逐步改进模型的预测性能。以下是一些常见的迭代最优化回归算法:

梯度下降法 (Gradient Descent): ① 批量梯度下降 (Batch Gradient Descent): 每次迭代使用整个训练集来计算梯度。② 随机梯度下降 (Stochastic Gradient Descent, SGD): 每次迭代使用一个样本来计算梯度, 更新参数。③ 小批量梯度下降 (Mini-batch Gradient Descent): 每次迭代使用一小部分样本来计算梯度。

动量法 (Momentum): 在梯度下降基础上引入动量项, 累积之前的梯度信息, 加速收敛并减少震荡。

AdaGrad: 自适应学习率算法, 根据历史梯度调整每个参数的学习率, 适合处理稀疏数据。

RMSProp: 改进的 AdaGrad 算法, 使用指数加权移动平均来计算梯度平方和, 适合处理非平稳目标。

Adam: 结合了动量法和 RMSProp 的优点, 自适应学习率和动量项, 广泛应用于深度学习。

这些算法通过不断迭代, 逐步逼近最优解, 适用于各种回归问题和复杂的机器学习模型。

本预测模型中最贴切的方法是梯度下降法 (Gradient Descent), 流程说明如下:

① 设定初始化参数: 学习率 $\alpha = \tau \cdot i$, 最大迭代次数 $m = 1/i$, τ 是学习率步长, 初始设定为 0.01, i 从 1 到 m 。

② 计算损失函数: 定义损失函数 $J(\theta)$, 如均方误差 (RMSE):

$$J(\theta) = f(RMSE, \alpha) \quad (11)$$

③ 计算损失参数的梯度: 计算损失函数对参数 α 的梯度:

$$\Delta J(\theta) = \frac{1}{m} \sum_{i=1}^m (f(RMSE, \alpha)_i - f(RMSE, \alpha)_{(i+1)})^2 \quad (12)$$

④ 迭代: 重复步骤 2 到 3, 直到达到损失函数的梯度收敛到 0 或者肘部出现。

7 预测范围的问题

在电力负载预测中, 如果预测值不是一个精确值而是一个上下波动的范围, 也是非常有价值的, 提高模型鲁棒性, 使其在面对实际数据波动时表现更好。可以帮助决策者制定更灵活、安全的计划, 加强了不确定性管理。电力需求和供应受多种因素影响, 如天气、经济活动和突发事件, 通过预测范围, 电力公司可以评估不同情景下的风险。例如, 如果预测范围较宽, 可能需要准备更多的备用电源以应对潜在的高需求; 如果预测范围收窄, 则可以资源优化避免过度或不足的准备。这些对范围的预测支持有关电力市场参与者在电力交易、价格设定和合同谈判中掌握更多的决策主动权。

在一个区域内稳定电力市场中, 电力负荷的变化率通

常会受到季节、天气、经济活动等多种因素的影响。根据统计数据,电力负荷的日变化率在不同季节和不同地区会有所不同。例如,在夏季和冬季,由于空调和取暖设备的使用,电力负荷通常会达到高峰,而在春秋季节则相对较低。具体到每天的波动率,通常在5%~10%波动。

布林线(Bollinger Bands)是由约翰·布林格(John Bollinger)发明的一种技术分析工具,最适宜应用于波动性、趋势性、周期性研究,尤其是它提供的上下限统计阈值,非常适合电力负荷的波动范围预测。当前布林线最成功最主流的应用领域是金融行业,将其在电力行业借鉴应用笔者尚未发现先例。布林线的计算方法:

中轨线(MID):通常是某一时间段(如20天)的移动平均线(MA)。

$$MA = \frac{1}{N} \sum_{i=1}^N V_{y(0)(i-1)} \quad (13)$$

$$SD = \sqrt{\frac{1}{N} \sum_{i=1}^N (V_{y(0)(i-1)} - MA)^2} \quad (14)$$

上轨线(UPPER):中轨线加上该时间段内分析值的标准差的两倍。

$$Upper\ Band = MA + 2 \times SD \quad (15)$$

下轨线(LOWER):中轨线减去该时间段内分析值的标准差的两倍。

$$Lower\ Band = MA - 2 \times SD \quad (16)$$

从数据研究角度进行验证,提取波动性、趋势性、周期性同样突出的A股上市公司中国神华(SH.601088)在2019-2020两年全部488个交易日数据,期间平均日波动2.6%,进行布林运算。结果显示,向上突破布林上轨的异常数据率为8.81%,向下突破为0%,通过回归测算,4倍SD宽度是平衡上下轨宽度和异常数据率的最佳取值,91.9%的区间包容率在实践中非常可行,这对类似统计波动率的电力负载区间预测具有重大参考意义。

2019—2020中国神华日成交量分布见图1。

8 结语

电力负荷预测不仅影响电力系统的供需平衡,也关系到供电可靠性,产业上下游资源优化,全社会综合节能减排效能。论文创新性提出一种基于数据挖掘的预测方法,对短期负荷预测以及短期预测区间分布进行研究,并阐述了与模式相关的各种算法的优越性,为有关方面的进一步细化实践提供了方向。

论文提出的预测模式还有多方面的拓展空间,如在时序的微观和宏观周期方面、在电力负荷数据特征数据集的向量维度方面;另外,如果将不同区域电网的预测结果进行实时链接、比较、复合可作为大电网调度的决策参考信息,在技术和经济方面体现更大价值。这些留待未来进一步进行探讨。

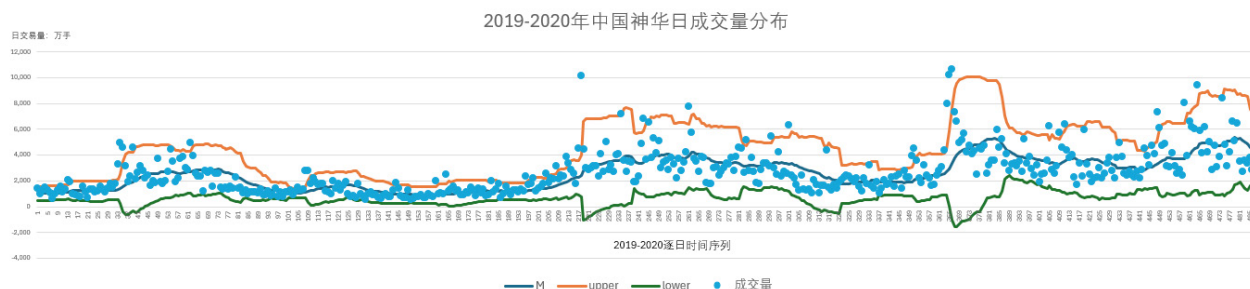


图1 2019—2020中国神华日成交量分布

参考文献

- [1] 潘攀,陈凡.基于数据挖掘的短期电力负荷预测方法[J].科学技术创新,2023(21):201-204.
- [2] 庞传军,张波,余建明.基于LSTM循环神经网络的短期电力负荷预测[J].电力工程技术,2021(1):175-180.
- [3] 韩雅萱,石梦舒,黄元生,等.基于机器学习的短期电力负荷预测方法比较及改进研究[J].科技管理研究,2023(1):163-170.
- [4] 苏伟,肖小龙,史明明,等.周期规律增强的多视角短期电力负荷预测[J].北京航空航天大学学报,2024(1):477-486.