

计及数据不平衡的 RUSBoost-LightGBM 短期负荷预测方法

张毅 温蜜

(上海电力大学计算机科学与技术学院 上海 201306)

摘要:随着人工智能技术的发展,负荷预测的准确度显著提高,但仍有一些负荷曲线预测精度较低。造成这种现象的原因在于训练集中这类负荷曲线属于少数类,导致数据不平衡,模型无法充分学习,从而影响了预测的精度。为解决电力负荷数据不平衡问题导致的预测精度降低的问题,提出了一种先分类后预测的解决方法。通过改进的 K-Medoids 方法把具有高相似性的历史负荷曲线进行聚类,并构造分类标签与电力负荷的特征集;然后根据预测时间的特征进行分类,通过 RUSBoost 算法很好的优化了分类过程中数据不平衡问题;最后在每类中使用 LightGBM 算法进行负荷预测。在公开数据集上的实验表明,所提出方法对少数类负荷的预测效果显著,其平均绝对百分比误差(MAPE)为 2.95%,均方根误差(RMSE)为 175.71 MW;对于常规负荷的预测,MAPE 为 3.52%,RMSE 为 195.84 MW,相对其他方法也具有较好的预测效果。

关键词:负荷预测;负荷曲线聚类;数据不平衡;RUSBoost;LightGBM

中图分类号: TM714 **文献标识码:** A **国家标准学科分类代码:** 470.40

Short-term load forecasting method with RUSBoost-LightGBM considering data imbalance

Zhang Yi Wen Mi

(College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306, China)

Abstract: The advancement in artificial intelligence technology has greatly enhanced load forecasting precision. Nonetheless, certain load curves within the training set experience low forecasting accuracy. This is attributed to data imbalance resulting from the load curves belonging to a limited number of classes. Consequently, the model fails to learn sufficiently leading to compromised forecasting accuracy. To address the issue of reduced prediction accuracy due to data imbalance in electric load forecasting, a solution involving classification before load prediction is proposed. The method involves clustering historically similar load profiles using an improved K-Medoids approach, constructing classification labels, and feature sets for electric load. Subsequently, classification is performed based on temporal features, and the RUSBoost algorithm is employed to effectively address data imbalance issues during the classification process. Finally, the LightGBM algorithm is utilized for load prediction within each class. Experiments on public datasets demonstrate the effectiveness of the proposed method in forecasting certain types of loads. The MAPE is 2.95%, and the RMSE is 175.71 MW. For the forecasting of conventional loads, the MAPE is 3.52%, and the RMSE metric is 195.84 MW, which is relative to other methods that also have better forecasting effect.

Keywords: load forecasting; load profile clustering; data imbalance; RUSBoost; LightGBM

0 引言

电力供需平衡是确保电网安全可靠运行的基础,而负荷预测则是分析电力供需的关键。准确的电力负荷预测

不仅有助于优化电网调度控制、确保电力系统安全稳定运行,还能提高经济和社会效益,保障人民正常生活和社会正常生产^[1]。通常,电力负荷预测依据时间长短和运行决策分为超短期、短期、中期和长期预测^[2]。短期负荷预测

是电网负荷预测的重要组成部分之一。它涉及预测未来1 h至2周内的电力负荷。通过短期预测电力负荷,可以提前评估该地区用电需求,从而合理规划发电量和调整供电方案,避免能源浪费或电力供应不足的情况。同时,也能避免负荷异常对电力设备造成的损害。

短期负荷预测目前的方法主要分为统计学方法与人工智能方法两大类。经典的统计学算法,如自回归滑动平均模型^[3]和差分移动平均自回归模型^[4],曾广泛用于电力负荷预测。然而,随着负荷变化日益复杂,统计方法的预测精度显著降低。相对而言,人工智能方法更能有效捕捉负荷变化的复杂规律。在机器学习领域,支持向量回归(support vector regression, SVR)^[5]、随机森林^[6]以及利用集成学习的XGBoost^[7]和LightGBM^[8]等方法已在短期负荷预测中取得了显著的成功。随着深度学习的兴起,BP神经网络^[9]、长短期记忆(long short term memory, LSTM)网络及其变体^[10-12]、门控递归单元(gated recurrent units, GRU)^[13],以及更为先进的Transformer模型^[14]也被广泛应用于短期负荷预测。

随着电网规模的扩大和智能电表的发展,负荷数据的规模不断扩大,在面对海量的历史负荷数据,直接进行训练会带来较大的误差和时间消耗。为解决此问题很多研究对负荷数据进行了聚类的划分。文献[15]提出了一种将住宅负荷配置文件聚类 and 索引为双级别负荷形状字典的方法,提升了聚类的性能和计算的效率,更好地捕获用户之间的常见用电模式。文献[16]通过改进的K-means聚类分析,用来区分工业用户不同生产模式的变化,识别不同地区和行业的负荷特性,提高了负荷预测的精度。文献[17]提出了模糊c均值聚类与动态时间规整(dynamic time warping, DTW)相结合的方法,对负荷数据进行聚类分析,并通过消融实验证明了对负荷预测的有效性。

负荷数据通过聚类的划分减少了模型的训练量,但聚类后的负荷数据存在着明显的数据不平衡问题,这会对少数类负荷的识别与预测带来误差。文献[18]提出一种计及类别平衡的两阶段选择集成电力负荷模式识别方法解决了集成学习负荷模式识别中的数据不平衡问题,有效减少了数据不平衡问题导致的少数类负荷样本辨识度被多数类淹没的机率。文献[19]提出了一种考虑样本不平衡的并行化用户负荷类型辨识方法,采用一种考虑样本合成倍率的改进Borderline-SMOTE方法处理负荷训练样本类别不平衡现象,有效提高小类负荷样本的分类精度。文献[20]考虑了不平衡小类样本导致负荷预测精度不高问题,采用基于概率分布合成小类样本方法扩充小类样本数据以平衡样本类别,提高了对节假日的少数类负荷预测精度。

以上研究对大量历史负荷序列进行了聚类分析,以识别少数类负荷曲线并考虑数据不平衡问题。然而,存在一些问题。首先,在众多研究中,为提高训练样本数量,通常将历史负荷曲线人工划分为几个主要类别。这种方法存

在分类不准确的缺陷,未考虑到少数异常负荷曲线的特殊性。将这些异常曲线纳入某一主类中进行训练,这可能导致少数类负荷预测中产生显著误差。其次,聚类完成后,对每一类进行训练集和验证集的划分,并检验模型的有效性。但这在现实的负荷预测中,会存在一些问题。未来的负荷曲线是未知的,无法提前进行聚类,并且无法简单地通过经验将其划分到特定类别。

基于上述问题,本文提出了一种先分类后预测的负荷预测框架。在该框架中,通过改进的K-Medoids算法针对电力负荷曲线的周期特性,对具有高度相似性的历史负荷曲线进行聚类,同时构建每一类负荷曲线的特征和标签。为解决未来负荷曲线分类匹配中的数据不平衡问题,引入了RUSBoost算法,采用欠采样技术平衡数据集,并通过集成学习方式解决了由于欠采样导致的信息丢失问题。最后,在分类后进行负荷预测时,本文采用了LightGBM算法。该模型利用基于梯度的单边采样算法、互斥特征捆绑算法和直方图算法,有效解决了大规模数据预测的挑战。与其他预测方法相比,所提出方法表现出了更好的预测效果。

1 算法理论介绍

1.1 改进的K-Medoids聚类

K-Medoids聚类是在经典的K-Means聚类上改进的,K-Medoids聚类是将集群中位于中心的点作为参考点,而不是K-Means中将集群中的平均值作为参考点^[21]。K-Means聚类对于样本中的异常值较为敏感,而改进后的K-Medoids减缓噪声数据造成极大的误差。K-Means++是一种改进的K-Means聚类算法,旨在更好地初始化聚类中心,以加速算法的收敛速度和提高聚类结果的质量^[22]。传统的K-Means算法在选择初始聚类中心时,通常是随机选择数据集中的 k 个点作为初始中心,导致聚类结果不稳定,还可能导致算法在某些情况下陷入局部最优解。

改进的K-Medoids聚类是将K-Means++的思路应用于K-Medoids聚类的方法,通过改进聚类初始点和聚类中心点的选择,缓解了K-Means聚类的两大缺陷,增强了聚类的鲁棒性和效果。改进的K-Medoids聚类算法步骤如下。

假设有一个包含 N 个数据点的电力负荷数据集,每个负荷数据点的维度为 D ,目标将这些数据点分成 K 个簇。首先,使用K-Means++的思想选择初始的 K 个中心点。令 C 代表当前已选择的中心点的集合, C_i 代表第 i 个中心点。

1) 选择第1个中心点,在数据集中随机选择一个数据点。

2) 选择后续中心点,对于每个未选择的数据点 x_j ,计算其与已选择中心点的最短距离(中位数距离),记为 $d(x_j, C)$ 。然后,按照式(1)选择下一个中心点 C_i 。

$$P(x_j) = \frac{d(x_j, C)^2}{\sum_{i=1}^N d(x_i, C)^2} \quad (1)$$

式中:分母是所有未选择的点的距离之和。通过按照这个概率选择下一个中心点,确保了距离较远的点更有可能被选择。

3)重复步骤2),直到选择了 K 个中心点。

有了初始中心点后,应用 K-Medoids 算法的迭代步骤。

分配样本到最近的中心点:对于每个数据点,计算其到所有中心点的距离 x_j ,将其分配到距离最近的中心点所在的簇。

$$c(x_j) = \operatorname{argmin}_{i=1}^K d(x_j, C_i) \quad (2)$$

更新中心点:对于每个簇 c_k ,计算簇内所有点对中位数的总体代价,选择具有最小总体代价的点作为新的中心点。

$$C_i = \operatorname{argmin}_{x_j \in c_k} \sum_{x_k \in c_k} d(x_k, x_j) \quad (3)$$

重复分配和更新的步骤,直到中心点不再发生变化或达到预定的迭代次数。

1.2 RUSBoost 算法

针对样本数据存在的严重的数据不平衡问题,本文利用 RUSBoost 算法来构建对少数类负荷曲线的分类预测模型,是一种欠采样技术与集成学习方法的结合。该算法利用欠采样的方法平衡数据集,缩短了模型训练的时间。同时,在每次迭代的基学习器训练集中都是独立采样,因此该算法可以有效克服因欠采样而造成的信息丢失问题。已有研究也表明,相较于其他采样算法而言,RUSBoost 算法是一种更加高效的方法,可以更好地处理数据不平衡问题^[23]。

假设 $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, 其中, S 为输入的训练集, $x_i \in X$ 表示第 i 个样本的特征向量, $y_i \in Y = \{-1, +1\}$ 表示第 i 个样本的类别标签。此外,弱分类器为 WeakLearn;最大迭代次数 T ,输出为分类器 $H(x)$ 。RUSBoost 的算法步骤如下

1)初始化训练集权重 $D_1(i) = 1/m, i = 1, 2, \dots, m$ 。

2)对于 $t = 1 \sim T$, 执行如下步骤。

(1)使用权重 D_t 训练一个弱分类器 $h_t(x)$, 其中 $h_t: X \rightarrow \{-1, +1\}$ 。

(2)计算分类误差率:

$$\epsilon_t = \sum_{i=1}^m D_t(i) [h_t(x_i) \neq y_i] \quad (4)$$

(3)计算弱分类器的权重:

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t} \quad (5)$$

(4)更新样本权重:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \quad (6)$$

其中 Z_t 是规范化因子,使得 D_{t+1} 成为一个概率分布,即:

$$Z_t = \sum_{i=1}^m D_t(i) \exp(-\alpha_t y_i h_t(x_i)) \quad (7)$$

3)输出分类器

$$H(x) = \operatorname{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right) \quad (8)$$

式中:RUSBoost 算法使用了 RUS(random under-sampling)技术来平衡类别分布。在每次迭代中,RUSBoost 会从少数类中随机选择一些样本,将其从训练集中删除,以达到类别平衡的目的。

1.3 LightGBM 算法

LightGBM 是微软研究院 2017 年提出的基于决策树算法的梯度提升框架,广泛用于分类或回归任务^[24]。LightGBM 与 XGBoost 算法都是基于 GBDT 算法的分布式梯度提升框架,但与 XGBoost 算法相比有着更快的训练速度和更少的内存消耗,并且在许多场景下有着更好的准确率。LightGBM 算法做了如下改进。

1)基于梯度的单边采样算法,通过对样本采样的方法来减少计算目标函数增益时候的复杂度,从而来缩小分裂点的搜索区域。

2)互斥特征捆绑算法,高维特征的分布往往存在稀疏的情况,特征间可能是相互排斥的,但如果两个特征并不完全互斥,可以用互斥率来表示互斥程度。该算法指出如果将一些特征进行融合绑定,则可以降低特征数量。

3)直方图算法,将连续的特征离散化为 k 个离散特征,并构造一个宽度为 k 的直方图用于统计信息(含有 k 个 bin)。利用直方图算法无需遍历数据,只需要遍历 k 个 bin 即可找到最佳分裂点。该算法有效减少内存与时间的消耗。直方图的构建过程如图 1 所示。

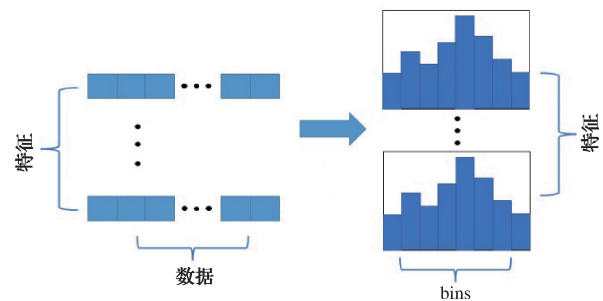


图 1 直方图构建过程

Fig. 1 Histogram construction process

2 提出计及数据不平衡的 RUSBoost-LightGBM 短期负荷预测方法

整体预测流程如图 2 所示,该负荷预测分为两个主要部分,分类框架和预测框架。在分类的基础上,预测过程考虑了负荷数据的不平衡情况,以提高对少数类的负荷预测精度。同时,对负荷数据进行了更准确的分类,为后续

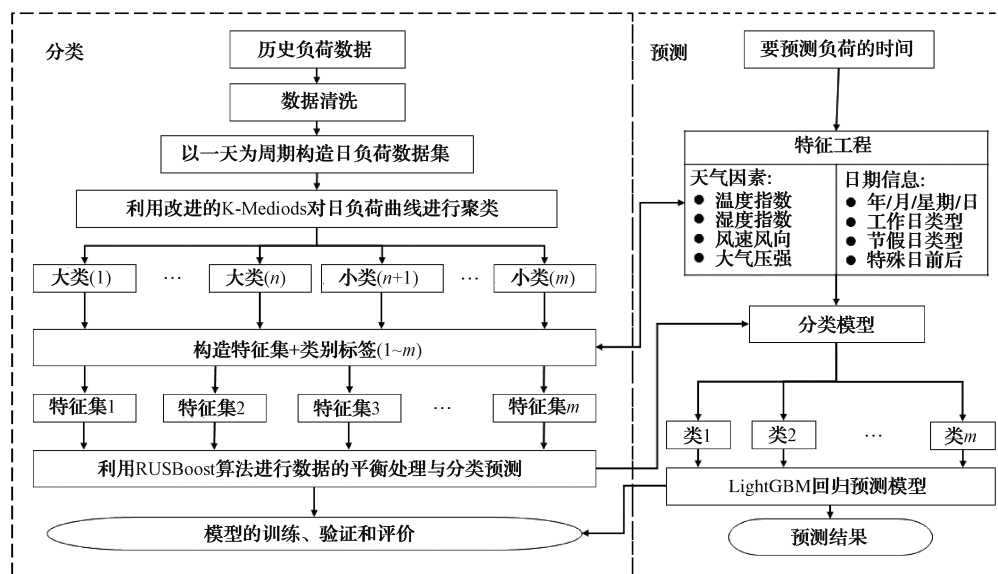


图2 提出预测方法流程

Fig. 2 Flowchart of the proposed prediction method

模型训练提供了更具针对性的数据子集,从而增强了预测模型的效能。

本文采用了数据清洗与重构的方法对数据集进行处理,首先按照特定的时间尺度对数据进行划分,随后进行分类与预测。通过调整输入数据的时间间隔和模型的时间窗口大小,可以灵活地改变预测时间的范围。因此,本文的模型能够有效地实现对未来1 h至2周内短期电力负荷的预测。

2.1 分类模型

在分类模型中,原始的负荷数据集通常受到缺失值和异常值的影响,对后续的预测产生显著影响,因此需要对数据进行预处理。本文采用箱线图识别异常值,并通过线性插值的方法填补缺失值,以确保数据的质量和完整性。负荷曲线的变化与人类生产活动息息相关,都具有明显的周期性,最小的周期单元是1 d。故本文以1 d的时间尺度进行划分,使用改进的K-Medoids算法进行聚类。该算法利用K-Means++的思路优化了K-Medoids算法中初始聚类中心的选择,并通过计算轮廓系数(silhouette coefficient, SC)确定聚类的个数 m 。改进的K-Medoids算法流程如图3所示。

在聚类完成后,每一类负荷曲线都对应着一个特定的标签,在考虑预测时间特征的情况下,为每一类负荷构造特征集。这样便构成了特征加标签的特征集,然而,这导致了明显的样本不平衡问题。为了更清晰地解释此问题,本文通过二维的不平衡数据分布图进行说明。如图4所示,在少数类样本和多数类样本比例为1:19的情况下,即使是简单的线性模型,将所有样本划分为多数类,准确率却高达到了95%。显然,这样的模型和评价方式是不合适的。为解决这一问题,本文采用了RUSBoost算法来

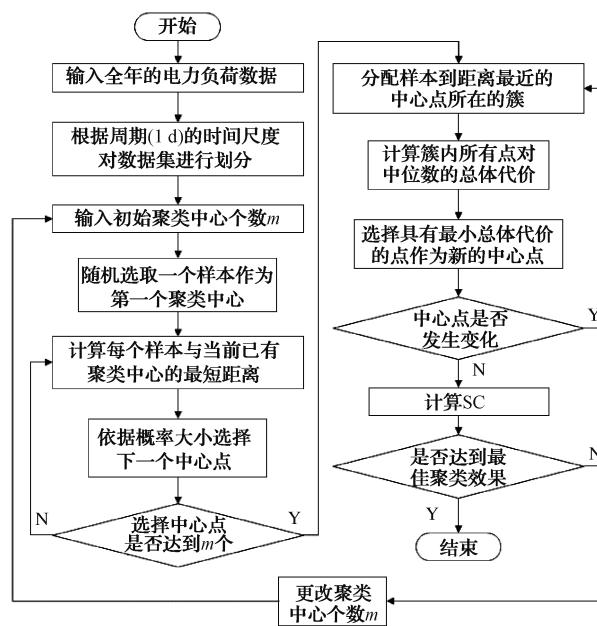


图3 改进的K-Medoids算法流程

Fig. 3 Flowchart of the improved K-Medoids algorithm

提高分类器对少数类别的识别能力,并使用更加合理的分类指标进行评估。

2.2 预测模型

在负荷预测模型中,短期负荷预测所需的时间特征是已知的。对于少数类负荷的出现,通常表现为该时间段特征的相对特殊性。为了实现对少数类负荷的精准预测,在特征工程阶段需要综合考虑多种因素。在天气因素方面,通过准确的天气预报可有效获取未来一天的温度、湿度、风速等特征;在日期因素方面,基于日历信息,未来时间的

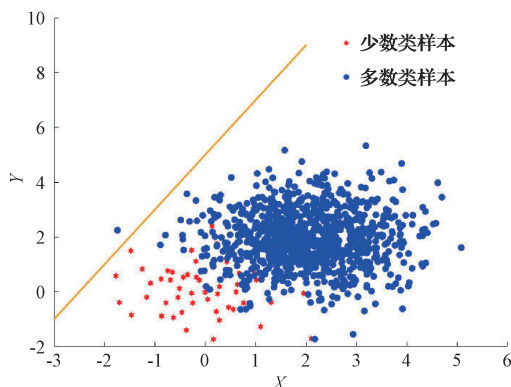


图4 不平衡数据分布

Fig. 4 Distribution of imbalance data

工作日类型、节假日类型、是否为特殊日等特征也是已知的。输入特征及其编码方式如表1所示。

表1 输入特征及编码

Table 1 Input features and coding

数据类型	特征	编码方式
时间相关数据	月类型	离散(1~12)
	日类型	离散(1~31)
	工作日类型	离散(0~1)
	节假日类型	离散(0~2)
天气相关数据	最高温度	连续
	最低温度	连续
	平均温度	连续
	相对湿度	连续
	降雨量	连续

基于预测时间的相关特征,将其输入到经过训练的分类模型中,分类完成后使用 LightGBM 算法对每个类别进行回归预测。通过在整个框架中进行分类和预测未来时间,成功实现了负荷预测。这一方法不仅考虑了数据不平衡情况下的少数类负荷,还对常规负荷进行了更为精准的预测。

3 数据集与评估指标

3.1 数据集

数据集来自2016年“电工杯”数学建模竞赛。其中包含某地区从2012年~2013年的电力负荷以及气象数据。选取2012全年的负荷数据来训练提出的模型。为了检验该模型对少数类特殊负荷的预测性能,另选取特殊日(元旦)的24个时间点作为测试集。

对2012全年的日负荷数据进行了分析如图5所示,横坐标单位为1d的时间(每15min记录一次,1d共96个数据),纵坐标为日期(全年366d),竖坐标为负荷量。总体上每天的负荷变化有着一定的相似性,但具体分析当中依然存在着不同的变化趋势。

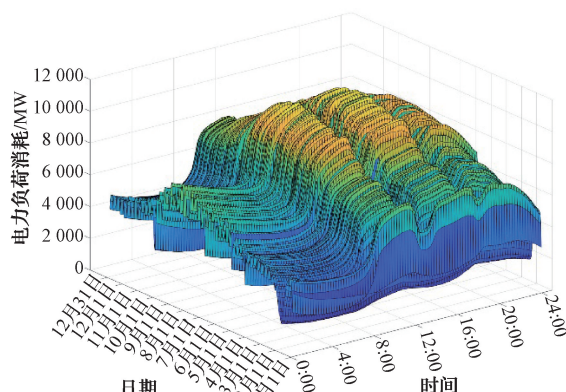


图5 2012年全年负荷分布

Fig. 5 Load distribution in 2012

3.2 评估指标

本文方法包含分类与预测的两部分,指标也相应的分为两个部分,一个是分类效果的指标,另一个是回归预测的指标。此外,在分类模型中为证明所提出改进聚类方法的有效性,增加了聚类指标进行评估。

1) 聚类指标

在负荷序列聚类的效果评估上,实验使用轮廓系数对日负荷聚类进行评估。轮廓系数是通过衡量样本与其自身聚类内部的紧密度和与其他聚类间的分离度进行评估。取值范围在 $[-1, +1]$,值越大聚类效果越好。轮廓系数公式如下:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (9)$$

式中: $a(i)$ 为样本 i 到同簇其他点的平均距离(簇内平均距离); $b(i)$ 为样本 i 到最近簇中的其他样本的平均距离(簇间平均距离)。

2) 分类指标

在分类结果评价指标中,由于数据不平衡的原因,本文在考虑总体分类准确率(Accuracy)的基础上增加了对指定正类的真正率(true positive rate, TPR)和 F_1 分数。

采用 Accuracy、TPR、 F_1 分数作为类别不平衡数据集分类的评价指标,分类效果需根据表2所示的二分类混淆矩阵进行计算与评估,对少数类进行分类评估时,要将正类设置为少数类,负类为多数类。

表2 混淆矩阵

Table 2 Confusion matrix

分类	预测为正类 (少数类)	预测为负类 (多数类)
	TP	FN
真实为正类(小类别)	TP	FN
真实为负类(多数类)	FP	TN

Accuracy 表示总体分类正确的概率,公式如下:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (10)$$

TPR表示指定正类预测的精确率,假负率(false negative rate,FNR)则表示指定正类预测的错误率,公式如下:

$$\begin{cases} TPR = \frac{TP}{TP + FN} \\ FNR = \frac{FN}{TP + FN} \end{cases} \tag{11}$$

正预测值(positive predictive value,PPV)表示预测指定正类的召回率,假发现率(false discovery rate,FDR)则表示预测指定正类的错误召回率,公式如下:

$$\begin{cases} PPV = \frac{TP}{TP + FP} \\ FDR = \frac{FP}{TP + FP} \end{cases} \tag{12}$$

F_1 是衡量精确率和召回率的分类指标,常用于评价不平衡数据集的分类预测,公式如下:

$$F_1 = \frac{2 \times TPR \times PPV}{TPR + PPV} \tag{13}$$

2)预测指标

在预测结果的评价指标中,采用平均绝对百分比误差(mean absolute percentage error,MAPE)和均方根误差(root mean square error,RMSE)两个指标来评估模型的预测能力。MAPE是实际负荷和预测负荷的绝对误差之和除以实际值的平均值,是评价预测模型可靠性的最常用指标。RMSE用于测量预测负荷值的总体偏差,它与电力负荷具有相同的单位,这使得评估模型的预测能力变得方便和直接。MAPE和RMSE定义为:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - y'_i|}{y_i} \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |y_i - y'_i|^2} \tag{15}$$

式中: n 是预测负荷点的数量; y_i 和 y'_i 分别代表实际负荷值和预测负荷值。

4 实验效果与分析

4.1 聚类效果分析

对2012全年日负荷曲线进行聚类分析,为证明所提出聚类方法的优势,使用轮廓系数对聚类结果进行评价,并与其他聚类方式进行了比较。各聚类算法在聚类数为6时,聚类效果达到了最优,轮廓系数值最高。与本文对比的算法有K-Means,K-Means++和K-Medoids算法,聚类评价结果如表3所示。

表3 聚类评价

Table 3 Clustering evaluation

方法	指标 SC
K-Means	0.563
K-Means++	0.571
K-Medoids	0.589
改进的 K-Medoids	0.611

由表3可知,改进的K-Medoids算法聚类效果最好,轮廓系数SC指标为0.611,明显优于其他聚类算法。利用改进的K-Medoids聚类后,各类别日负荷曲线数目如表4所示。

表4 聚类结果

Table 4 Clustering results

聚类类别	数目
1	86
2	49
3	16
4	133
5	68
6	14
总计	366

在分类预测中,当不平衡比例超过4:1时,通常会导致对多数类的偏向预测。值得注意的是,上述聚类结果显示有明显的数据不平衡现象,其中最少样本数量的类别(类别6)与总样本数的比例约为1:26。此外,类别2和类别3的样本数量占总样本数的比例也小于1/7,因此它们也属于少数类别,存在分类预测中可能出现误差的风险。因此,在后续的分类工作中,必须考虑由数据不平衡问题引起的预测精度下降。

4.2 分类效果分析

1)对比实验

RUSBoost算法是欠采样算法与AdaBoost算法相结合的模型。欠采样解决了数据不平衡带来的问题,为证明其有效性,将RUSBoost算法与AdaBoost算法进行了比较,同时也与其他分类算法进行了比较。各类算法的总体分类准确率如表5所示。

表5 准确率结果

Table 5 Accuracy results (%)

方法	RUSBoost	AdaBoost	随机森林	SVM	决策树
Accuracy	82.5	81.7	76.9	77.4	77.1

RUSBoost与AdaBoost在总体的分类准确率上差别较小,且因为集成学习的原因高于其他算法。考虑数据不平衡问题,将类别数量最少的类别6设置为正类,分别计算TPR和 F_1 值,结果如表6所示。

表6 各模型分类结果评价

Table 6 Evaluation of classification results of each model

LightGBM 预测	MAPE/%	RMSE/MW
未做数据分类	7.74	745.25
分类但未平衡数据集(AdaBoost)	4.57	447.45
分类且平衡了数据集(RUSBoost)	3.52	195.84

结果表明,使用 RUSBoost 分类算法处理后的 TPR 值和 F_1 值明显高于其他算法,说明相对与 AdaBoost 及其他算法对少数类的正确分类能力有了明显的提升,证明了对数据不平衡问题处理的有效性。

2) 消融实验

以最后负荷预测结果为标准,对数据集的分类处理和平衡处理做消融实验。对未作分类的负荷数据集进行直接预测,同时与使用分类模型的训练结果进行对比。在分类模型中,未经过平衡数据的分类模型是 AdaBoost 算法,使用平衡数据集技术的分类模型是 RUSBoost 算法。其中经过分类的结果取所有类别的平均误差。实验结果如表 7 所示。

表 7 对分类与平衡数据集技术的消融实验结果

Table 7 Results of ablation experiments on categorized and balanced dataset techniques

方法	RUSBoost	AdaBoost	随机森林	SVM
TPR/%	84.1	63.4	53.9	41.5
F_1	0.700 1	0.554 2	0.568 2	0.352 0

实验结果表明,直接使用 LightGBM 算法对未经分类处理的数据集进行预测,其结果明显低于经过分类处理的

方法。在已分类的算法中,从最终预测结果看出,对数据进行平衡处理的预测指标仍然优于未进行平衡处理的方法。这可能是因为未平衡数据集的算法对测试集中的少数类产生了较大的误差。

4.3 预测效果分析

为了证明本文所提出方法对特殊和正常情况下负荷预测的有效性,模型在 2012 年全年数据上训练,训练集上使用 5 折交叉验证检验训练效果,另选取 2013 年 1 月 1 日作为测试集。为证明选取测试集的特殊性,将该天的负荷与其前后 10 d 的日负荷数据做了皮尔逊相关性分析,如图 6 所示,结果表明 1 月 1 日的负荷与其前后 10 d 的日负荷有着明显的不同,属于本文所属的少数类负荷。根据预测时间的特征经 RUSBoost 算法分类后,属于其中的第 6 类(少数类)负荷曲线,也验证了其负荷的特殊。为验证本文提出的方法在少数类负荷预测方面的效果,将其与其他算法进行了比较,各模型的预测结果如表 8 所示。对比的算法是时间序列预测中知名且有效的算法,其中包括 LSSVM、TCN、LSTM、Bi-GRU、LSTM-CNN、LSTM-CNN-Attention、Transformer、Informer 模型。鉴于测试集涉及特殊日,为确保模型的公正性,在对比算法中还加入了相同的特征。

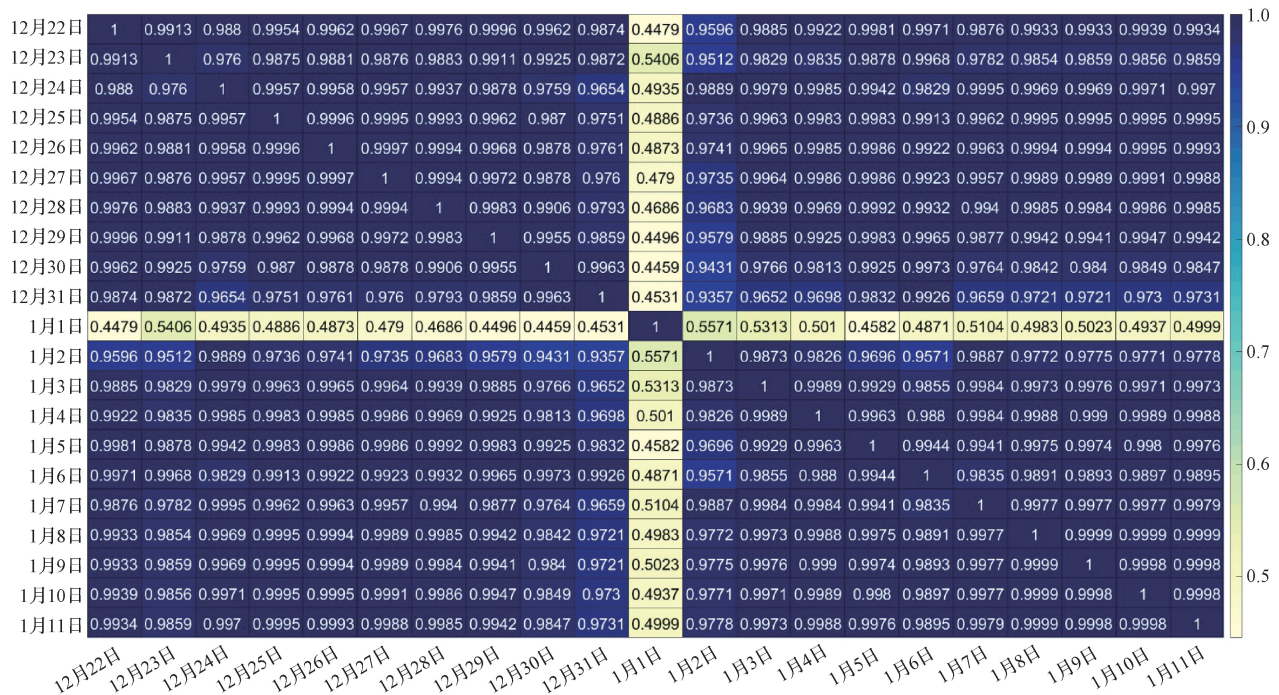


图 6 日负荷皮尔逊相关热力图

Fig. 6 Daily load Pearson correlation heat map

本文提出的预测方法相对于其他方法对于少数类负荷情况(测试集)有着明显的预测精度提升,并且对全年所有的负荷(训练集)预测依然有着不错的预测效果。值得注意的是在训练集上本文提出的方法在 RMSE 的评估上

优势更为明显,原因是 RMSE 在计算中对误差进行了平方求和,对大误差惩罚更大,这也证实了本文方法在对特殊负荷的预测上具有更好的鲁棒性。虽然该预测框架在对少数类负荷预测有着明显优势,但这并不影响对正常

表 8 各模型预测结果数据
Table 8 Data on the forecasting results of each model

时间	2012 全年 (训练集)		2013 年 1 月 1 日 (测试集)	
	MAPE /%	RMSE /MW	MAPE /%	RMSE /MW
LSSVM	7.54	765.27	41.03	1764.23
TCN	4.09	336.95	3.80	389.85
LSTM	3.67	304.88	4.93	408.63
Bi-GRU	3.54	365.27	4.85	394.93
LSTM-CNN	3.84	258.34	3.71	341.96
LSTM-CNN-Attention	3.92	248.13	3.73	369.36
Transformer	3.98	394.65	4.03	415.23
Informer	5.17	456.08	11.05	902.26
本文方法	3.52	195.84	2.95	175.71

负荷的预测精度,反而因为类别的细分提高了预测精度。

5 结 论

本文针对电力负荷数据不平衡问题导致的预测精度降低的问题,提出了一种先分类后预测的解决方法。对历史负荷曲线的处理上,通过改进的 K-Medoids 算法对负荷数据进行了更准确的聚类,为后续模型训练提供了更具有针对性的数据子集。在分类过程中,特别考虑了数据不平衡问题,通过引入 RUSBoost 算法显著提升了对于少数类负荷的分类准确率。最后,使用 LightGBM 算法进行回归预测,生成负荷预测值。为验证所提出方法的有效性,进行了消融实验与对比实验,结果表明关键步骤在整体预测效果上起到了显著提升作用。与其他先进算法相比,本文提出的模型框架在特殊负荷(少数类)预测方面表现出显著的效果提升,同时对正常负荷的预测依然有着较高的精度。

参 考 文 献

[1] 夏博,杨超,李冲. 电力系统短期负荷预测方法研究综述[J]. 电力大数据,2018,21(7):22-28.
XIA B,YANG CH,LI CH. Review of the short-term load forecasting methods of electric power system[J]. Power Systems and Big Data,2018,21(7):22-28.

[2] TANVEER A,ZHANG H C,YAN B. A review on renewable energy and electricity requirement forecasting models for smart grid and buildings[J]. Sustainable Cities and Society 2020,55: 102052.

[3] BOROOJENI K G,AMINI M H,BAHRAMI S, et al. A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon[J]. Electric Power Systems Research, 2017, 142: 58-73.

[4] LEE C M,KO C N. Short-term load forecasting

using lifting scheme and ARIMA models[J]. Expert Systems with Applications, 2011, 38(5): 5902-5911.

[5] 陈家瑞,陈忠孝,秦刚,等. 基于 PSO 算法与 SVR 算法在企业直流配电网短期负荷预测的研究[J]. 国外电子测量技术,2020,39(12):70-73.
CHEN J R, CHEN ZH X, QIN G, et al. Research on short-term load prediction of enterprise DC distribution network based on PSO algorithm and SVR algorithm[J]. Foreign Electronic Measurement Technology,2020,39(12):70-73.

[6] FAN G F,ZHANG L Z,YU M, et al. Applications of random forest in multivariable response surface for short-term load forecasting[J]. International Journal of Electrical Power & Energy Systems, 2022, 139: 108073.

[7] 常乐,汪庆年. 基于优化聚类分解与 XGBOOST 的超短期电力负荷预测[J]. 国外电子测量技术,2022, 41(5):46-51.
CHANG L, WANG Q N. Ultra-short-term power load forecasting based on optimal cluster decomposition and XGBOOST [J]. Electronic Measurement Technology,2022,41(5):46-51.

[8] 陈伟楠,胡志坚,岳菁鹏,等. 基于长短期记忆网络和 LightGBM 组合模型的短期负荷预测[J]. 电力系统自动化,2021,45(4):91-97.
CHEN W N, HU ZH J, YUE J P, et al. Short-term load prediction based on combined model of long short-term memory network and light gradient boosting machine[J]. Automation of Electric Power Systems, 2021,45(4):91-97.

[9] 张淑清,杨振宁,张立国,等. 基于弹性网降维及花授粉算法优化 BP 神经网络的短期电力负荷预测[J]. 仪器仪表学报,2019,40(7):47-54.
ZHANG SH Q, YANG ZH N, ZHANG L G, et al. Short-term power load forecast based on dimension reduction by elastic network and flower pollination algorithm optimized BP neural network[J]. Chinese Journal of Scientific Instrument, 2019,40(7):47-54.

[10] LIN J, MA J, ZHU J, et al. Short-term load forecasting based on LSTM networks considering attention mechanism [J]. International Journal of Electrical Power & Energy Systems, 2022, 137: 107818.

[11] 夏飞,张洁,张浩,等. 基于 BIC 准则和加权皮尔逊距离的居民负荷模式精细识别及预测[J]. 电子测量与仪器学报,2020,34(11):33-42.
XIA F, ZHANG J, ZHANG H, et al. Fine recognition and prediction of resident load pattern based on BIC criterion and weighted Pearson

- distance[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(11): 33-42.
- [12] 孟金鑫, 黄山, 印月. 基于特征优选策略和 DLSTMs-FCN 优化的短期负荷预测[J]. 电子测量技术, 2023, 46(10): 46-52.
- MENG J X, HUANG SH, YIN Y. Short-term load forecasting model based on feature optimization strategy and DLSTMs-FCN [J]. Electronic Measurement Technology, 2023, 46(10): 46-52.
- [13] 吴永洪, 张智斌. 基于贝叶斯优化的 CNN-GRU 短期电力负荷预测[J]. 现代电子技术, 2023, 46(20): 125-129.
- WU Y H, ZHANG ZH B. CNN-GRU short-term power load forecasting based on Bayesian optimization[J]. Modern Electronics Technique, 2023, 46(20): 125-129.
- [14] 石卓见, 冉启武, 徐福聪. 基于聚合二次模态分解及 Informer 的短期负荷预测[J/OL]. 电网技术, 1-15 [2024-05-11]. <https://doi.org/10.13335/j.1000-3673.pst.2023.1467>.
- SHI ZH J, RAN Q W, XU F C. Short-term load forecasting based on aggregated secondary decomposition and informer[J/OL]. Power System Technology, 1-15[2024-05-11]. <https://doi.org/10.13335/j.1000-3673.pst.2023.1467>.
- [15] LIANG H S, JIN M. Develop load shape dictionary through efficient clustering based on elastic dissimilarity measure [J]. IEEE Transactions on Smart Grid, 2020, 12(1): 442-452.
- [16] GE Q B, GUO C, JIANG H Y, et al. Industrial power load forecasting method based on reinforcement learning and PSO-LSSVM[J]. IEEE Transactions on Cybernetics, 2020 52(2): 1112-1124.
- [17] TANG X L, CHEN H X, XIANG W H, et al. Short-term load forecasting using channel and temporal attention based temporal convolutional network[J]. Electric Power Systems Research, 2022, 205: 107761.
- [18] 王圆圆, 韩丁, 王世谦, 等. 面向类别不平衡负荷序列模式识别的两阶段选择集成学习策略[J]. 电力系统及其自动化学报, 2023, 35(1): 86-95.
- WANG Y Y, HAN D, WANG SH Q, et al. Two-stage selective ensemble learning strategy enabling pattern recognition of class imbalanced load series[J]. Proceedings of the CSU-EPSA, 2023, 35(1): 86-95.
- [19] 刘洋, 高丽霞, 刘璐. 考虑样本不平衡的并行化用户负荷类型辨识方法[J]. 电网技术, 2020, 44(11): 4310-4317.
- LIU Y, GAO L X, LIU L. Parallel load type identification algorithm considering sample class imbalance [J]. Power System Technology, 2020, 44(11): 4310-4317.
- [20] 杨秋玉, 旷树森, 郑小刚, 等. 基于 NPMA-LSSVM 算法的不平衡小类样本情况下中短期负荷预测[J/OL]. 电测与仪表, 1-13 [2024-05-11]. <http://kns.cnki.net/kcms/detail/23.1202.TH.20220907.1549.002.html>.
- YANG Q Y, KUANG SH S, ZHENG X G, et al. Medium and short-term load forecasting based on NPMA-LSSVM algorithm in the case of unbalance and minority sample data [J/OL]. Electrical Measurement & Instrumentation, 1-13 [2024-05-11]. <http://kns.cnki.net/kcms/detail/23.1202.TH.20220907.1549.002.html>.
- [21] PARK H S, JUN C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36 (2): 3336-3341.
- [22] ARTHUR D, SERGEI V. K-means ++ the advantages of careful seeding[C]. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007.
- [23] SEIFFERT C, KHOSHGOFTAAR T M, HULSE J V, et al. RUSBoost: A hybrid approach to alleviating class imbalance[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2009, 40 (1): 185-197.
- [24] KE G L, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.

作者简介

张毅, 硕士研究生, 主要研究方向为电力负荷预测与综合能源负荷预测。

E-mail: zhangyi19981001@163.com

温蜜, 博士, 教授, 主要研究方向为智能电网与无线通信网络安全。

E-mail: miwen@shiep.edu.cn