DOI:10.20033/j.1003-7241.(2025)05-0099-05.

基于电力大数据挖掘的异常用电用户识别模型研究

郑 真，马晔晖，黄一楠

(国网上海市电力公司青浦供电公司,上海 201700)

摘 要: 为有效管理电网用电用户,保障电力系统运行安全,提出基于电力大数据挖掘的异常用电用户识别模型研究。采用大数据挖掘方法获取用户用电量、用电类型两种用电行为大数据,经归一化处理后,将处理后的用电行为大数据,输入至由长短期记忆网络和双向门控循环单元组建的深度循环神经网络中,挖掘用户用电行为的时序特征,通过逻辑回归模型分类用户用电行为的时序特征,实现异常用电用户识别。经实验验证,该模型能够有效分类识别正常与异常用电用户,通过用电量与日平均负荷判断用户是否存在异常用电。

关键词: 大数据挖掘;异常用电用户;识别模型;用户用电量;深度循环网络;时序特征

中图分类号:TP183;TM76 文献标识码:A 文章编号:1003-7241(2025)05-0099-05

Research on Abnormal Power User Identification Model Based on Power Big Data Mining

ZHENG Zhen, MA Ye-hui, HUANG Yi-nan

(State Grid Shanghai Qingpu Electric Power Supply Company, Shanghai 201700 China)

Abstract: In order to effectively manage the power users of the power grid and ensure the operation safety of the power system, a model for identifying abnormal power users based on power big data mining is proposed. The big data mining method is used to obtain the big data of consumer's electricity consumption and power consumption type. After normalization, the processed big data of power consumption behavior is input into the deep cycle spiritual network composed of long-term and short-term memory network and two-way gated cycle unit, mining the time sequence characteristics of users' power consumption behavior, and classifying the time sequence characteristics of users' power consumption behavior through the logical regression model. It realizes identification of abnormal power users. The experimental results show that the model can effectively classify and identify normal and abnormal electricity users, and judge whether there is abnormal electricity consumption by electricity consumption and daily average load.

Keywords: big data mining; abnormal power users; recognition model; consumer's electricity consumption; deep recurrent network; timing characteristics

0 引言

随着我国社会的高速发展,用电用户海量增多,当前我国电网供电线路损失率逐渐增高,线路损失的核心原因就是异常用电行为十分严重[1],当大量的异常用电现象发生,电网企业供电成本加剧,同时,由于难以追究异常用电的根源,导致这些行为对电网企业带来严重的危害[2]。因此,在信息化逐渐普及的现代社会,供电企业可利用电力运行数据对异常用电用户进行识别,提高异常用电行为的检测能力,才能够使电力企业保持稳定运行[3-4]。

海内外有较多学者对异常数据识别进行了研究,国外学者Hu X提出基于随机检测的工业物联网安全异常行为检测算法[5],该算法通过随机形式对异常数据进行了大规模的检测,但该算法仅适用于物联网领域,对于其他领域的异常数据检测与识别存在一定的缺陷。而国内有较多学者对电网数据的异常识别进行了研究,例如李清[6]研究电力大数据异常检测方法,通过大数据的聚类实现异常数据检测,虽然其检测速度较快,但该方法仅能够检测到电力系统内的异常数据内容,无法有效获取异常用电用户。万磊等[7]研究电力大数据用电异常检测方法,该方法利用长短期记忆网络(long short-term memory,LSTM)网络训练形式不断挖掘异常用电数据,但该方法的检测结果不够精准,导致部分用电行为存在遗漏。大数据挖掘技术是一种能够对海量数据同时进行处理的有效方式,将这一技术应用在电网异常用电用户识别中,可以迅速定位异常用户,强化数据的处理速度。为此,本文研究基于电力大数据挖掘的异常用电用户识别模型,获取异常用电用户,保障电力企业运行安全。

*基金项目:国网上海市电力公司科技项目(520934220005)

收稿日期:2024-01-26

1 电力企业异常用电用户识别模型

1.1 基于电力大数据挖掘的用户用电行为分析

通过电力大数据挖掘技术获取电网用户的用户用电量与用电类型，为电网异常用电用户识别提供有力依据。

1.1.1 用户用电量计算

电网用户在单位时间段内的电能消耗总量，即为用户用电量。在利用大数据挖掘获取用户用电信息时，每一用户均存在较大意义，因此，通过对每一用户的用电数据挖掘，可以得到他们的用户用电量情况[8]。根据用户用电量，可以体现出用户用电状态[9−10]。设电网用户使用电能的高频振动时长为$|T|$，用户最多消耗电流为I'，用户最多消耗电压为U'，根据这些物理量，可通过公式(1)计算用户用电量：

$$Q=\int_{T_0}^{T_1}\left[\frac{\overline{R}\left(\lambda\cdot|T|\right)}{\left(I'-I_0\right)\overline{R}}\right]^{-1}\left(U'-U_0\right)^2 d\overline{R} \tag{1}$$

式中，电能高频震动时长的最小、最大表现实值依次为T_0、T_1，用户电压最低消耗为U_0，用户电流最低消耗为I_0，在电网节点位置的电阻均值为$\overline{R}$，电量的既定输入系数为λ。

1.1.2 用户用电类型划分

用户用电类型划分是指对用电用户的具体用电内容进行详细化描述，其中包含电流、电压以及实用电量三项目标。实质上是指当用户用电内容发生改变时，电网的电压、电流以及电量在消耗过程中均会发生改变，同时改变趋势并不存在规律。在大多数情况下，当用户的用电情况逐渐明显，会使实际运行的电压与电流传输量上涨，并导致电量出现变化。

假设用户用电过程中的电压累积量为$\overline{U}$，电流累计量为$\overline{I}$，结合公式(1)，可通过公式(2)计算电网用电用户的电压、电流累计量：

$$\begin{cases}\overline{U}=\dfrac{1}{Q}\sum_{s}^{s\to+\infty}\max\left(\dfrac{\chi\cdot\overline{y}}{|\varpi_1-\varpi_0|}\right)\\ \overline{I}=\dfrac{Q\cdot\min\left(\dfrac{\mu\cdot\overline{y}}{|\chi_1-\chi_0|}\right)}{s^2}\end{cases} \tag{2}$$

式中，用电用户节点负载电阻的最小值为s；ϖ_0、ϖ_1均为干扰压降差，两者存在一定差距，干扰压降差之间的平均数为ϖ；χ_0、χ_1均为干扰流降差，两者的值同样存在区别，干扰流降差之间的平均数为χ，用户节点处电阻定值消耗参量为$\overline{y}$。

假设在整个电力大数据挖掘过程中，用电用户节点处的电能使用情况并未出现明显波动，此时，设传输电压利用率为ξ，并设传输电流利用率为η，结合公式(2)中用电用户电压电流累计量的计算，可通过公式(3)划分电网环境中用户的用电情况：

$$D=\sum_{p=1}l^{\gamma+1}\frac{\sqrt{\xi\overline{U}}}{\eta\overline{I}} \tag{3}$$

式中，D为用电用户使用的电能情况；标准用电量为l，电压与电流之间消耗限定参数为γ。

1.2 电网用电用户大数据预处理

由于在用电用户行为分析过程中可能会出现电表异常等现象[11−13]，因此需要对用电用户的用电大数据进行归一化处理，使用电用户大数据使用更加方便。通过用电大数据的归一化处理，不仅可以降低后续运算的复杂度[14]，还可以防止大数值属性支配小数值属性，通过公式(4)实现电网用电用户的大数据归一化：

$$x=\frac{x'-x_{mean}}{x_{\max}-x_{\min}} \tag{4}$$

式中，经归一化处理前、后的用电用户大数据集为x'、x；用电用户大数据样本中的最大值、最小值依次为$x_{\max}$、$x_{\min}$；用电用户大数据样本平均值为x_{mean}。

1.3 深度循环神经网络的用电行为时序特征提取

LSTM是一种循环神经网络，该网络适用于处理时间间隔较长的事项。双向门控循环单元(gated recurrent unit，GRU)是一种简化版的LSTM块，同样属于神经网络结构，通过这两种网络的结合，可有效处理梯度消失的问题，还能够精准学习时序的依赖特征，因此可以更完善地提取大数据的时序特征。本文将归一化后的电网用电用户大数据输入到LSTM块与GRU块相结合的深度循环神经网络中进行训练，获取用电用户大数据的时序特征，为用电用户的异常用电识别提供可靠依据。本文构建的LSTM网络由许多LSTM块组成，其中每个块均存在输入门、输出门与遗忘门，通过图1描述每个LSTM块结构的结构与训练过程。

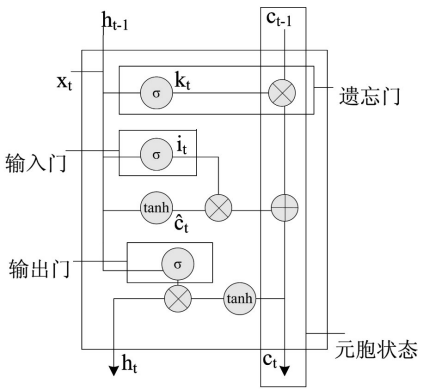

图1 LSTM块的结构与训练过程

通过图1的结构实现LSTM块的训练，其具体训练步骤如下：

(1) 将内容输入到sigmoid层中，即遗忘门，输入内容

为x_t、h_{t-1}，其中，x_t表示经归一化后t时刻下的用电用户大数据，h_{t-1}表示为$t-1$时刻下LSTM神经网络的输出。该遗忘门输出值处于0～1之间。同时设权重矩阵为W，偏置量为b，sigmoid函数为$\sigma(\cdot)$。假设遗忘门输出值为1，则表示该用电用户大数据保留，若输出值为0则丢弃该用电用户大数据，遗忘门输出通过公式(5)进行计算：

$$k_t=\sigma\left(W_k\cdot\left[h_{t-1},x_t\right]+b_k\right) \tag{5}$$

式中，k_t表示遗忘门输出，W_k为遗忘门权重，b_k为遗忘门偏置。

(2) 将步骤(1)输出的结果传输至输入门，在输入门中通过sigmoid函数决定对哪些用电用户大数据参数进行更新，并通过式(6)计算该输入门的输出：

$$i_t=\sigma\left(W_i\cdot\left[h_{t-1},x_t\right]+b_i\right) \tag{6}$$

式中，i_t表示输入门的输出，W_i表示输入门的权重，b_i表示输入门偏置。

(3) 在输入门中通过tanh函数构建新的候选向量$\hat{c}_t$，通过式(7)判定该向量是否添加到状态中：

$$\hat{c}_t=\tanh\left(W_c\cdot\left[h_{t-1},x_t\right]+b_c\right) \tag{7}$$

式中，W_c、b_c为候选向量$\hat{c}_t$的权重与偏置。

同时，通过式(8)，利用候选向量$\hat{c}_t$更新得到新的用电用户大数据状态单元：

$$c_t=k_t\cdot c_{t-1}+i_t\cdot\hat{c}_t \tag{8}$$

式中，c_t为新状态单元。

根据式(8)中状态单元的更新结果，将新状态单元c_t输入到输出门中，在输出门利用式(9)、式(10)输出状态单元c_t的最终形式：

$$o_t=\sigma\left(W_o\cdot\left[h_{t-1},x_t\right]+b_o\right) \tag{9}$$

$$h_t=o_t\cdot\tanh\left(c_t\right) \tag{10}$$

公式(9)通过输入一个sigmoid函数决定输出单元状态的哪一部分，其中，o_t表示输出门的输出内容，W_o表示输出门权重，b_o表示输出门偏置；h_t为t时刻下LTSM网络的输出。

本文通过GRU网络对LSTM网络中的状态单元进行优化，通过GRU单元可以有效获取大数据在不同时间尺度下的依赖性，相较于LSTM网络，GRU单元可以去除记忆单元，避免训练过程变得复杂，使模型训练速度加快。在GRU单元中，主要包含重置门与更新门，其中，重置门用于将新的信息与之前的记忆结合，通过式(11)计算对重置门进行计算：

$$r_t=\sigma\left(W_r\cdot\left[c_{t-1},x_t\right]\right) \tag{11}$$

式中，r_t表示重置门输出，W_r表示重置门的权重，c_{t-1}表示$t-1$时刻下的用电用户大数据状态单元。

通过更新门可以决定之前计算的记忆是否保留，还能够决定是否将新的内容传输至下一状态，通过式(12)表示更新门：

$$z_t=\sigma\left(W_z\cdot\left[c_{t-1},x_t\right]\right) \tag{12}$$

式中，z_t表示更新门输出，W_z表示更新门权重。

若之前的隐藏状态为无关的，则可通过更新门将该状态消除。通过更新门产生的用电用户大数据状态元素为新的记忆，该状态元素由过去隐藏状态和新输入决定，具体如式(13)所示：

$$\hat{c}_{t-1}=\tanh\left(W\cdot\left[r_t\cdot c_{t-1},x_t\right]\right) \tag{13}$$

式中，$\hat{c}_{t-1}$表示$t-1$时刻下的候选向量，r_t为隐藏状态。

当获取具备新的记忆的状态元素后，利用更新门的影响，通过之前的隐藏输入与新的记忆组合构成新的隐藏状态，如式(14)所示：

$$c_t=c_{t-1}\cdot\left(1-z_t\right)+z_t\cdot\hat{c}_{t-1} \tag{14}$$

式中，当前GRU网络的输出即为c_t，即为优化后的状态单元，c_{t-1}为表示前一个GRU网络的状态单元输出。

通过LTSM网络提取电网用电用户大数据的时序特征，并利用GRU神经网络优化LTSM网络参数，构成深度循环神经网络，经不断训练后，得到更为精准的电网用电用户大数据时序特征。

1.4 基于逻辑回归模型的异常用电用户分类识别

利用电网用户用电大数据时序特征，通过逻辑回归模型对电网异常用电用户进行分类识别。由于逻辑回归模型属于双分类结构，因此该模型的计算结果仅为1和0，将用电用户通过G→{0,1}的映射表示，其中1为异常用户，0为正常用户。为了能够利用之前提取到的电网用电用户大数据时序特征识别出用户用电类别，本文在逻辑回归模型中，引入一个Sigmoid函数，该函数计算公式如下：

$$h_\theta\left(\hat{x}\right)=\frac{1}{1+e^{-\theta^T\hat{x}}} \tag{15}$$

式中，h为LSTM网络的输出，$\hat{x}$为一个电网用电用户大数据时序特征组建的多维向量，$\hat{x}=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_i,\cdots,\hat{x}_n\}$，$\hat{x}_i$即为某一电网用电用户大数据时序特征，$\hat{x}_i$对应的特征参数为$\theta_i$，其中$\theta=\{\theta_1,\theta_2,\cdots,\theta_i,\cdots,\theta_n\}$，$\theta^T$表示$t$时刻下的电网用电用户大数据时序特征参数，$h_\theta$为包含用电特征的用电用户大数据。若$h_\theta(x)=1$，则说明识别到的电网用电用户为异常用户，否则为正常用户。

假设训练样本共有m个，每个上标i表示第i个样本，为使电网用电用户大数据分类识别的输出结果$\hat{h}^{(i)}$更接近真实结果$h^{(i)}$，即$\hat{h}^{(i)}\approx h^{(i)}$，本文定义$m$个训练样本的代价函数$J(\tau,\psi)$。通过式(16)计算电网用电用户识别的代价函数平均损失：

$$J(\tau,\psi)=\frac{1}{m}\sum_{i=1}^{m}L\left(\hat{h}^{(i)},h^{(i)}\right)=$$
$$-\frac{1}{m}\sum_{i=1}^{m}\left[\left(h^{(i)}\log\left(\hat{h}^{(i)}\right)+\left(1-h^{(i)}\right)\log\left(1-\hat{h}^{(i)}\right)\right)\right] \tag{16}$$

式中，L为似然函数。

通过式(16)可有效评估识别结果与真实结果之间的平均错误代价，对代价函数$J(\tau,\psi)$进行最小化，即可实现识别结果的优化。本文采用梯度下降法，对代价函数中的权重τ与偏移量ψ进行最小化优化，如下所示：

$$\tau=\tau-\alpha\frac{\alpha\left(J(\tau,\psi)\right)}{\partial\tau} \tag{17}$$

$$\psi=\psi-\alpha\frac{\alpha\left(J(\tau,\psi)\right)}{\partial\psi} \tag{18}$$

式中，α、∂表示梯度函数。

由于梯度朝着负方向移动，因此通过梯度下降法可代价函数的获取最小值，当权重τ与ψ为最小状态时，即可获取最小代价函数J，得到最佳代价函数，使异常用电用户识别的结果更靠近真实结果。

2 仿真实验

2.1 数据集来源

本文采用爱尔兰智能能源试验电表真实数据，对本文提出的识别模型进行验证，在该数据集中主要包含三种类型的用户，分别为居民、工业以及其他用户，其中每种用户均包含535天用电记录，在每条记录中包含一天48次用电信息采样点，实验从中选取1 000个用户信息进行分析，从中得到535×1 000=535 000条用电记录，在这些用电信息中，具体样本类型如表1所示。

表1 数据集用电类型

样本数量/条	居民用户用电	工业用户用电	其他用户用电
正常用户样本数/条	220 000	100 000	120 000
异常用户样本数/条	45 000	30 000	20 000

2.2 结果与分析

若居民用户日均用电量在3～5 kWh左右，工业用户日均用电量在25～27 kWh左右，假设某一居民用户在3月～4月份出现异常用电行为，而某一工业用户在4月～6月份同样存在异常用电行为，分析两者在不同月份下的用电量情况，以此评估所提模型的异常用电识别能力，分析结果如图2所示。根据图2可知，经过所提模型的识别后，可获取两用户的详细异常用电情况，当居民用户在3月～4月份出现异常用电时，该用户的用电量从4 kWh左右最高上升至12 kWh左右，说明该用户在此期间大量用电；而工业用户在4月～6月份出现异常用电时，工业用户的用电量大幅上升，最高达到36 kWh以上，因此，通过所提模型可有效识别得到异常用电用户的用电量变化，实现异常用电用户的精准识别。

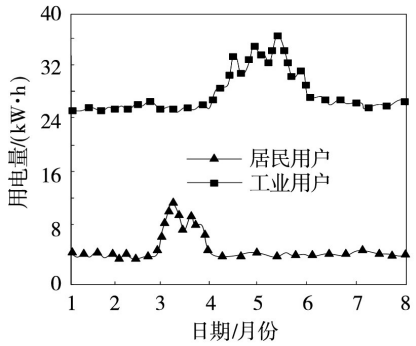

图2 用户用电行为分析

以用电用户的日平均负荷为评估标准，当用户存在异常用电行为时，其日平均负荷相对较大。采用所提模型对数据集中的三家居民用户进行用电行为识别，分析不同用户是否存在异常用电行为，分析结果如图3所示。根据图3可知，通过所提模型，可以获取三个居民用户在某日不同时间段的日平均负荷现象，三个用户的负荷均在7:00～16:00保持较高水平，说明在该时间段内居民用电较为集中，所用电量较大，但是，用户3的日平均负荷明显高于其他两用户，且该用户最高日平均负荷达到17 kWh左右，由此可以看出，用户3存在异常用电行为，通过所提模型的识别，可精准识别出居民用户的异常用电状态，以及在异常用电发生时的负荷变化。

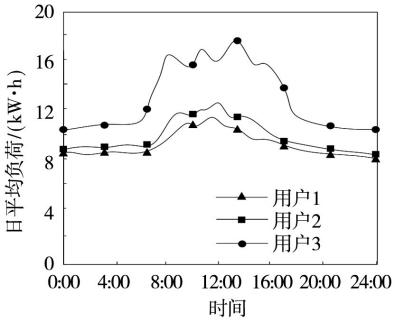

图3 用户日平均负荷变化情况分析

表2 所提模型识别结果

用户编号	是否存在异常用电	异常用电类型	异常特征	识别输出结果
居民用户1	是	欠流法窃电	电表输入电流异常	存在异常用电
居民用户2	是	移相法窃电	功率因数异常	存在异常用电
居民用户3	是	欠压法窃电	电表输入电压异常	存在异常用电
工业用户1	否	不存在窃电	-	正常用电
工业用户2	是	移相法窃电	功率因数异常	存在异常用电
工业用户3	是	欠流法窃电	电表输入电流异常	存在异常用电
其他用户1	否	不存在窃电	-	正常用电
其他用户2	是	移相法窃电	功率因数异常	存在异常用电
其他用户3	是	扩差法窃电	电能计量误差异常	存在异常用电

采用所提模型对数据集中的10个不同类型的用户进

(下转第154页)