文章编号:1671-4598(2023)10-0054-07 DOI:10.16526/j.cnki.11—4762/tp.2023.10.009 中图分类号:TP183 文献标识码:A

基于门控循环单元的非均衡数据驱动异常用电检测方法

孟宋萍，彭 伟，田晨璐
（山东建筑大学 信息与电气工程学院，济南 250101）

摘要：异常用电检测能够及时发现异常用电行为，在减少能源浪费和经济损失的同时能够维持安全、稳定的电网运行环境；智能电表的普及使得用电数据获取十分容易，为数据驱动的异常用电检测方法提供了充足的数据支持；然而，在实际应用过程中，异常数据较少导致的数据非均衡问题严重影响了模型的训练效果；因此，针对上述问题提出了一种针对非均衡数据的门控循环单元异常用电检测方法；该方法利用边界合成少数类过采样技术实现了对少数类数据的有效扩充；为了更好地捕捉用电数据的时序特征，采用了门控循环单元实现对用电数据的分类；为了验证该方法的有效性，基于非均衡数据集进行了对比实验；实验结果表明，该方法能够达到更好的数据扩充效果以及更准确的异常用电检测效果。

关键词：异常用电检测；异常用电行为；数据非均衡；边界合成少数类过采样；门控循环单元；时序特征

Imbalanced Data-driven Abnormal Electricity Consumption Detection Method Based on Gated Recurrent Units

MENG Songping, PENG Wei, TIAN Chenlu
(School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China)

Abstract: Abnormal electricity consumption detection can detect abnormal electricity consumption behaviors in time and maintain a safe and stable power grid operating environment while reducing energy waste and economic losses. The popularity of smart meters makes it very easy to obtain electricity consumption data, which provides sufficient data support for data-driven abnormal electricity consumption detection methods. However, the problem of imbalanced data seriously affects the training effect of the model in practical application. Therefore, in this paper, a gated recurrent units based abnormal detection method for imbalanced electricity consumption data is proposed. The method adopts the borderline synthetic minority oversampling technique to realize the effective extension of the minority data. To better capture the time series characteristics of electricity consumption data, gated recurrent units are employed to classify electricity consumption data. To verify the effectiveness of this method, comparative experiments are done on imbalanced data. Experimental results show that this method has better data expansion effect and more accurate abnormal electricity consumption detection effect.

Keywords: abnormal electricity consumption detection; abnormal electricity consumption behaviors; imbalanced data; borderline synthetic minority oversampling; gated recurrent units; time series characteristics

0 引言

随着人们生产、生活对电能的依赖性增强，对于其质量与可靠性的需求也在增长。而国内的能源结构及分布制约了我国相关行业的发展[1]。为了应对发展中面临的问题，大力发展智能电网成为了其中的解决方案之一。

智能电网的发展，使得用电过程中的问题尤其是异常用电问题暴露出来。异常用电作为一种非法行为，一直受到相关部门的控制。但是随着智能电网的发展，异常用电的技术手段越来越多，越来越不易被发现，异常用电的检测问题日益严重。

在美国，每年因异常用电损失 60 亿美元[2]，而我国每年损失大概 200 亿元[3]。异常用电行为在带来损失的同时也给电网的安全、稳定的运行带来了一定难度[4]。智能电表的普及，一方面阻止了某些异常用电行为的发生[5]，另一

收稿日期:2023-01-10； **修回日期**:2023-02-20。
基金项目:山东省重大科技创新工程(2021CXGC011205)；山东省科技型中小企业创新能力提升工程(2021TSGC1053)。
作者简介:孟宋萍(1997-),女,硕士。
通讯作者:彭 伟(1986-),男,博士,副教授,硕士生导师。
引用格式:孟宋萍,彭 伟,田晨璐.基于门控循环单元的非均衡数据驱动异常用电检测方法[J].计算机测量与控制,2023,31(10):54-60.

方面提供了大量的用电数据用于分析检测，一定程度上降低了异常用电造成的损失。但是目前异常用电所造成的能源浪费在经济损失上仍占很大的比例，对于异常用电的检测方法也存在一定的提升空间。

随着智能电表的普及，大量的用电数据为数据驱动的异常用电检测方法提供了数据支持。数据驱动的异常用电检测方法主要可以分为基于聚类、基于回归以及基于分类的三类。其中，回归和分类属于有监督学习方法，聚类属于无监督学习方法。

基于聚类的异常用电检测方法是将相似的用电数据通过特定算法划分成一个类别。文献［6］通过最优路径森林聚类方法实现对异常用电的检测，并且与 k-均值聚类和高斯混合模型等聚类方法进行了对比，验证了该方法的优越性。文献［7］采用了模糊 C-均值聚类来检测用户中的异常用电行为，并且可以根据模糊程度来判断其异常的程度。基于聚类的异常用电检测方法好处是不需要带标签的数据即可实现异常用电检测。但是，其缺点是聚类方法对参数的依赖性较高，参数选取通常比较困难。

基于回归的异常用电检测方法是根据历史用电数据以及各类用电影响因素对未来用电量进行预测，再根据预测量与实际用电量对比来确定是否存在异常用电行为。文献［8］使用了差分整合移动平均自回归模型和神经网络对天然气的用量进行了预测并且判断是否存在异常。文献［9］中的作者采用基于线性回归的方法来确定单个房屋的异常，并从房屋数据中清除此类异常，从而提供能源消耗模式的精确评估。但是，在实际生活中，用户的用电量与各种因素相关比如温度，天气状况等，并且随机性较强，因此很难依靠基于回归的方法实现较高精度的检测。

基于分类的异常用电检测方法可以将其分为机器学习方法和深度学习方法。经典的机器学习方法在异常用电检测中发挥了重要作用。文献［10－11］中，作者提出了基于 K-近邻（KNN，K-nearest neighbor）的算法来检测异常用电。文献［12－13］中，作者使用支持向量机来诊断由窃电而导致的异常。文献［14］中，作者改进了决策树模型，利用异常类和正常类的密度来检测消费数据中的异常。集成方法也为异常用电检测贡献了力量。文献［15］中，作者提出了梯度树增强（GBT，gradient boosting tree）方法来检测异常用电行为。文献［16］中，作者提出了以随机森林作为分类器的模型来检测异常用电。

随着深度学习进入大众的视野，基于深度学习的方法也被成功应用于异常用电检测中。在文献［17］中，作者设计了一种基于循环神经网络的异常检测系统，该系统可以从数据中去除季节性因素，从而能更好地捕捉数据的真实分布。文献［18］中，作者使用循环神经网络和 K-均值的混合模型识别异常消费。文献［19－20］中，作者提出了基于自动编码器和长短期记忆网络的方法识别用电数据中的异常。文献［21］中，作者提出了变分循环自编码器来检测异常。文献［22］中，作者将随机森林与卷积神经网络结合来检测窃电行为。而在文献［23－24］中，作者提出了基于卷积神经网络的模型，并且将用电数据转成二维数据来学习数据特征。

尽管异常用电检测已经取得了很多成果，但是仍然存在着很多问题。其中最重要的问题就是用电数据存在严重的非均衡性。因为用电数据涉及到用户的隐私，所以用户一般不会公开其用电数据。即便公开，可以得到的也是正常的用电数据，异常数据几乎没有。如果数据集中正常数据的数量远远大于异常数据数量，那么在训练检测模型时，模型更倾向于学习正常数据，不能学到异常数据的数据特征，导致检测效果较差。

合成少数类过采样技术的广泛应用为解决该问题提供了思路。合成少数类过采样技术通过线性插值合成新样本，实现少数类样本和多数类样本数量的均衡。文献［25］中，作者使用合成少数类过采样技术生成岩石可灌浆性分类数据。文献［26］中，作者使用合成少数类过采样技术扩充冷水机组故障数据。因此，在本文，可以借助上述思想，使用边界合成少数类过采样技术（BSMOTE，borderline synthetic minority oversampling technique）对异常数据进行扩充，得到数据平衡的数据集，然后再用于异常用电的检测中。

另外，由于用电数据是典型的时间序列数据，因此如何选择分类器也是一个重要问题。门控循环单元（GRU，gated recurrent units）是循环神经网络的变体，通过其内部的门结构可以实现对时间序列数据长期特性的记忆，并且可以缓解梯度消失的问题。文献［27］中，作者使用门控循环单元解决时间序列中长时间依赖问题用于手势识别。在文献［28］中，作者使用门控循环单元用于语音识别。受上述工作的启发，在本文，使用 GRU 作为用电数据的分类器，实现对异常用电的检测。

为了解决上述非均衡数据以及时间序列特性问题，提出了基于门控循环单元和边界合成少数类过采样技术的异常用电检测方法（GRU-BSMOTE），本文的贡献及创新点如下。

1）使用 BSOMTE 解决数据非均衡问题。使用 BSMOTE 对实现对少数类异常数据的有效扩充，使其数量与正常数据保持一致。该过程能够有效缓解因异常数据不足导致的模型训练不佳的问题。

2）为了更好地捕获用电数据的时间序列特征，使用 GRU 对用电数据进行分类。GRU 能够有效学习数据的时间特征，在减少训练时间的情况下解决长时间依赖和梯度消失的问题。

3）为了验证该方法的有效性，基于非均衡数据集做了详细的对比实验。实验结果表明，该方法能够实现在不同扩充比例情况下对数据的有效扩充，并且能以更高的准确率实现对异常用电的检测。

1 基本方法介绍

1.1 边界合成少数类过采样技术

在实际应用中，常见的数据非均衡问题的解决方法有 3 种，分别是数据过采样、欠采样和模型算法的改进。欠采样是指少数类样本数量不影响模型训练的情况下，对多数样本欠采样，实现样本数据的均衡。过采样是指少数类样本数量不足以支持模型的训练时，对少数类样本过采样，使其与多数类样本数量保持一致。模型算法的改进主要是提升模型对于少数类样本的学习能力。基于上述方法综合考虑后，在本文使用过采样技术对异常用电数据进行扩充。

在各种过采样方法中，合成少数类过采样技术（SMOTE，synthetic minority oversampling technique）是一种常用的方法，通过合成少数类样本来均衡数据集中各类样本的分布，提高非均衡数据集的分类精度。合成少数类过采样技术的原理是在相距较近的少数类样本之间生成新样本，没有充分考虑近邻样本的分布特点，存在一定的盲目性，非常容易造成数据类别之间的重复。而位于边界中的样本又对于模型进行分类决策有着重要作用。因此，本文使用边界合成少数类过采样技术对数据进行处理，实现对于非均衡数据集分类精度的提升。边界合成少数类过采样技术是在少数类样本的边界样本中合成新样本，可以有效避免上述问题的发生，提高生成新样本的质量，提高模型学习各类样本特征的能力，其原理如图 1 所示，并且详细介绍了其步骤。

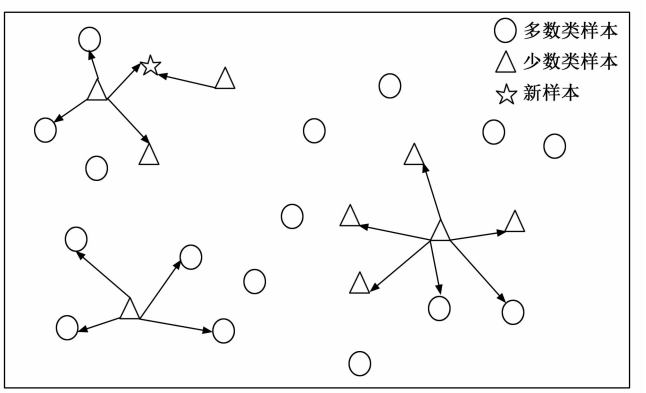

图 1 边界合成少数类过采样技术原理图

步骤 1：计算少数类样本的每个样本点 p_i 与所有样本的欧式距离，得到该样本的 m 近邻。

步骤 2：对少数类样本进行划分。设在 m 近邻中有 m' 个多数类样本，显然 $0 \leqslant m' \leqslant m$。如果 $m' = m$，那么 p_i 是噪声；如果 $\frac{m}{2} \leqslant m' < m$，那么 p_i 是边界样本；如果 $0 \leqslant m' < \frac{m}{2}$，那么 p_i 是安全样本。将边界样本记为 $\{p'_1, p'_2, \ldots, p'_i, \ldots, p'_{dnum}\}$，其中 $dnum$ 是少数类样本中边界样本的个数。

步骤 3：计算边界样本点 p'_i 与少数类样本 P 的 k 近邻，根据采样倍率 U，选择 s 个 k 近邻与 p'_i 进行线性插值，合成的少数类样本为 $s_j = p'_i + r_i \times d_j, j = 1, 2, \ldots, s$，其中 d_j 表示 p'_i 与其 s 个 k 近邻的距离，r_i 是 0 与 1 之间的随机数。

由于数据各个类别的边界数据对于模型的训练分类效果有着重要的作用，因此，边界合成少数类过采样技术在边界样本中合成新样本，合成的少数类新样本的分布更加合理，更加有利于模型区分各类数据，实现分类准确率及精度的提高。

1.2 门控循环单元

长短期记忆网络（LSTM，long short-term memory）作为特殊的循环神经网络，主要是为了解决长时间依赖以及梯度消失等问题。长短期记忆网络拥有 3 个由 Sigmoid 和点积操作构成的门结构，通过 3 个门结构的配合实现对时间序列中信息的丢弃和保留。虽然长短期记忆网络对于长期记忆问题非常有效，但是因为其引入了很多内容，导致其参数变多，使得训练过程难度加大。

门控循环单元是将长短期记忆网络简化改进后的处理时间序列数据的模型。门控循环单元同样能解决长时间依赖以及梯度消失的问题，并且与长短期记忆网络不同的是，门控循环单元只有两个门结构，在输出时也取消了二阶非线性函数。在保证学习效果的基础上，门控循环单元可以有效减少训练时间。在本文，使用门控循环单元作为分类器实现对用电数据的分类。门控循环单元的原理如图 2 所示，并且详细介绍了其原理。

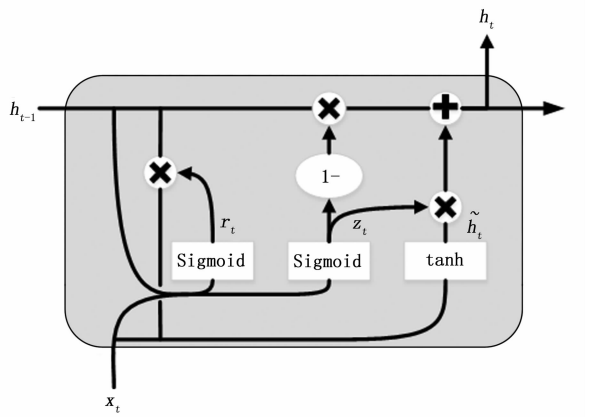

图 2 门控循环单元原理图

如图 2 所示，门控循环单元中的门结构都是由点积操作和 Sigmoid 构成，通过二者的配合可以实现对信息的丢弃和保留。门控循环单元的两个门结构分别是重置门和更新门。

首先，重置门 r_t 可以表示为：

$$r_t = \text{Sigmoid}(x_t \boldsymbol{W}_{xr} + h_{t-1} \boldsymbol{W}_{hr} + \boldsymbol{b}_r) \tag{1}$$

其中：x_t 是输入，h_{t-1} 是上一节点的隐藏状态，$\boldsymbol{W}_{xr}$ 和 $\boldsymbol{W}_{hr}$ 是权重矩阵，$\boldsymbol{b}_r$ 是偏置。Sigmoid 的取值是 0～1，因此可以充当门控信号，决定丢弃多少信息保留多少信息。

然后，更新门 z_t 可以写做：

$$z_t = \mathrm{Sigmoid}(x_t \boldsymbol{W}_{xz} + h_{t-1}\boldsymbol{W}_{hz} + \boldsymbol{b}_z) \quad (2)$$

得到上述的 r_t 和 z_t 后，候选隐藏状态 $\tilde{h}_t$ 可以表示为：

$$\tilde{h}_t = \tanh(x_t \boldsymbol{W}_{hx} + r_t \odot h_{t-1}\boldsymbol{W}_{hh} + \boldsymbol{b}_h) \quad (3)$$

其中：h_{t-1} 包含了过去的信息，r_t 是重置门，$\odot$ 是按元素相乘。tanh 激活函数可以将数据缩放到−1～1 的范围内。

最后，最终的隐藏状态 h_t 可以表示为：

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (4)$$

其中：z_t 的取值是 0～1，当 z_t 趋于 1 时，表示长期依赖一直存在。当 z_t 趋于 0 时，表示忘记隐藏信息中的不重要信息。门控循环单元的关键在于使用了同一个门控 z_t 即可实现对信息的遗忘和选择记忆。

总之，门控循环单元中的重置门决定了如何将当前输入信息与前面的记忆信息结合，更新门决定了前面的记忆有多少保存到当前时间。通过上述操作，可以解决对时间序列数据长期依赖问题，并且可以缓解梯度消失。

2　基于门控循环单元的非均衡数据异常用电检测方法

异常用电检测中的数据非均衡问题是指数据集中异常用电数据数量远远小于正常数据。在模型训练时，很难根据少量的异常数据学习到其特征，也就是说模型很难对异常数据进行检测识别，导致异常用电检测的效率低。

智能电表收集到的用户用电数据是典型的时间序列数据，选择怎样的模型对其进行分类尤为重要。循环神经网络是常用于时间序列数据分类或者预测问题的模型。虽然循环神经网络处理时序数据具有一定优势，但是它却无法解决时间序列中长时间依赖关系的问题，并且存在严重的梯度消失问题。

在本文，为了缓解非均衡数据导致的模型训练不佳的问题，使用 BSMOTE 对少数类数据进行扩充，得到平衡的数据集对模型进行训练。然后，为了更好的发掘时间序列数据的特性，解决时间序列中长期记忆以及梯度消失的问题，使用 GRU 构建用电数据与用电行为的映射关系。该方法的整体框架如图 3 所示，下面介绍了该方法的详细步骤。

步骤 1：对数据进行清洗，去除其中的异常值并且对使用平均值来代替其中的缺失值。

步骤 2：由于用电数据存在严重的非均衡问题，即正常用电数据的数量远远大于异常用电数据，使用 BSMOTE 对少数类数据进行扩充，得到平衡数据集。

步骤 3：将平衡数据集划分为训练数据集和测试数据集。使用训练数据集对门控循环单元进行训练、更新模型参数。测试数据集用于验证模型的训练效果。

值得注意的是，由于对异常用电检测模型训练使用的是由 BSMOTE 与真实数据构成的训练数据集，在测试时，一方面需要测试模型对于异常检测的准确率，另一方面也需要测试 BSMOTE 合成的数据是否可以用于异常用电检测模型的训练。因此，测试集数据应该全部是由真实数据构成，不仅可以测试模型的性能，还能够测试合成数据是否符合真实用电数据特性。

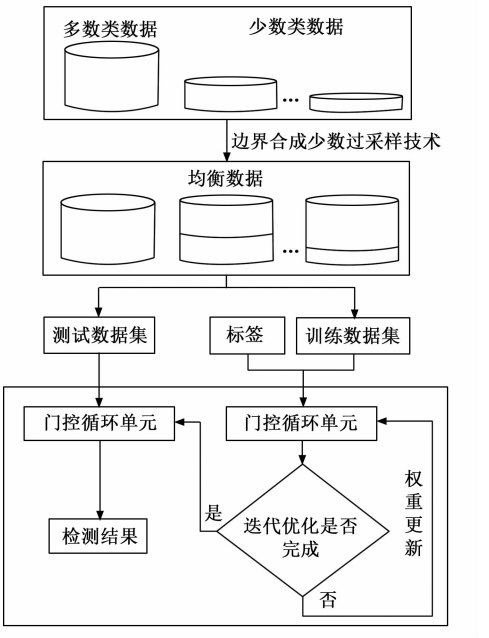

图 3　非均衡数据异常用电检测流程图

3　实验

3.1　实验数据和评价指标

在本文使用的数据集来自文献［29］，该数据集来自国外一家省级电力公司，其中包括了正常用电数据以及五类异常用电数据。在数据集中随机选取正常以及五类异常数据将其绘制在图 4 中。

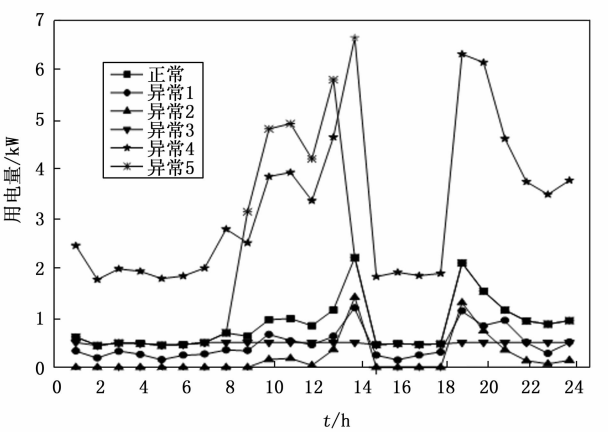

图 4　数据展示

如图 4 所示，正方形点所在的线代表了正常用电数据，其余 5 个线条代表了五类异常数据。其中，异常 1 表示用电量异常减少；异常 2 代表用户的主线路发生故障；异常 3 代表用户的支路线路发生故障；异常 4 代表用户用电量异常增加；异常 5 代表用户用电量在任意时间内异常增加。

另外，为了衡量模型应对非均衡数据的能力，使用了准确率（Acc，accuracy），精确度（P，precision），召回率

(R，recall)，和 $F1$ 分数 (F1，F1-score) 4 个指标。

准确率是预测正确的样本数量占总样本数量的比值，其公式如下：

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

其中：TP 代表样本实际是正类，模型将其预测为正类。TN 代表样本实际是负类，模型将其预测为负类。FP 代表样本实际是负类，但是模型却将其预测为正类。FN 代表样本实际是正类，但是模型将其预测为负类。

精确度是指所有预测为正类的样本中，实际也为正类的概率，计算公式为：

$$P = \frac{TP}{TP + FP} \tag{6}$$

召回率是指实际为正类样本，预测结果也是正类的概率，计算公式为：

$$R = \frac{TP}{TP + FN} \tag{7}$$

在应用中，精确度和召回率都希望很高，但是实际上二者是存在矛盾的，无法做到二者都最高，因此为了衡量二者的平衡，定义了 $F1$ 分数。$F1$ 分数可以同时考虑精确度和召回率，也就是说精确度和召回率的平衡点是 $F1$ 分数，其计算公式为：

$$F1 = \frac{2P \times R}{P + R} \tag{8}$$

3.2 对比方法及实验设置

在本文，将门控循环单元与经典分类模型支持向量机 (SVM，support vector machine) 以及时间序列模型长短期记忆网络做了对比。

SVM 作为典型的分类模型在故障诊断[30]和功率预测[31]方面取得了成功应用。SVM 通过寻找最优分类面实现对数据的分类。不仅可以对线性数据进行分类，借助核技巧将非线性数据映射到高维空间，使得 SVM 也可以处理非线性数据。

为了解决循环神经网络的无法学习到长期依赖以及梯度消失问题，LSTM 被提出[32]。LSTM 的优点是其拥有 3 个门结构，分别为遗忘门，输入门和输出门。每个门结构都是由一个 Sigmoid 层和点积操作组成。通过 3 个门结构的组合可以决定信息被保留多少和被丢弃多少。

本文搭建了循环层为 2 的堆叠 GRU 用于构建用电数据与用电行为的映射关系，其中隐藏层节点数为 32，损失函数设置为交叉熵损失函数，优化器设置为 Adam。在对比实验中，构建了一个双向 LSTM 模型，隐藏层节点数设置为 72。在使用非线性多维支持向量分类器对用电数据进行分类时，惩罚系数设置为 1，核函数设置为高斯径向基函数 (RBF，radial basis function)，参数 gamma 设置为 ‘auto’。

本文的所有实验都是在一台标准 PC 机上使用 Python 3.7 实现的，CPU 为 Intel 酷睿 i7-7700HQ，运行频率为 2.80 GHz，内存为 16.0 GB。

3.3 实验结果分析

3.3.1 验证 BSMOTE 的有效性

为了验证 BSMOTE 生成数据是否与真实数据相似可以用于模型的训练，使用生成数据作为训练集，真实数据作为测试集做了对比实验。另外，为了验证均衡数据集有利于模型的训练，还将扩充后的均衡数据集与非均衡数据集做了对比，并且考虑了不同数量真实数据的情况下即不同扩充比例的情况下，扩充后的均衡数据集的表现。扩充比例是指训练数据集中生成数据与真实数据的比值。实验结果如表 1 所示。

表 1 不同训练数据集异常用电检测结果

扩充比例	数据集	准确率/%	精确度/%	召回率/%	F1 分数
11∶1	非均衡	89.69	82.24	72.79	77.23
	均衡	99.07	99.09	99.07	99.08
5∶1	非均衡	89.71	88.66	85.73	87.17
	均衡	98.27	98.27	98.27	98.27
3∶1	非均衡	90.46	90.66	89.66	90.15
	均衡	97.97	97.98	97.97	97.98
2∶1	非均衡	91.29	91.72	92.13	91.93
	均衡	98.04	98.06	98.04	98.05
1∶1	非均衡	92.17	92.87	92.69	92.78
	均衡	97.55	97.59	97.55	97.57

从表 1 可以看出，当测试数据是真实数据时，异常用电的检测结果较好。当扩充比例为 11∶1 时，4 个指标均在 99%以上；当扩充比例为 5∶1 时，4 个指标均为 98.27%；当扩充比例为 3∶1 时，检测准确率为 97.97%；当扩充比例为 2∶1 时，4 个指标均在 98%以上；当扩充比例为 1∶1 时，异常用电检测精确度为 98.59%。上述数据说明使用 BSMOTE 生成的数据与真实数据是非常相似的，BSMOTE 在异常用电数据的扩充上是成功的。

另外，也可以看出不论生成数据与真实数据的比值是多少，与非均衡数据集相比，均衡数据集效果优于非均衡数据集。

详细来讲，在扩充比例为 11∶1 时，与非均衡数据集相比，准确率提高了 9.38%，精确度提高了 16.85%，召回率提高了 26.28%，F1 分数提高了 21.85%。在扩充比例为 5∶1 时，与非均衡数据集相比，准确率提高了 8.56%，精确度提高了 9.61%，召回率提高了 12.54%，F1 分数提高了 11.10%。在扩充比例为 3∶1 时，与非均衡数据集相比，准确率提高了 7.51%，精确度提高了 7.32%，召回率提高了 8.31%，F1 分数提高了 7.83%。在扩充比例为 2∶1 时，与非均衡数据集相比，准确率提高了 6.75%，精确度提高了 6.34%，召回率提高了 5.91%，F1 分数提高了 6.12%。在扩充比例为 1∶1 时，与非均衡数据集相比，准确率提高了 5.38%，精确度提高了 4.72%，召回率提高了 4.86%，F1 分数提高了 4.79%。

上述数据说明均衡的数据更有助于模型的训练，有助

于模型容易学习到不同类别数据的特征，提高模型的分类精度。

3.3.2　数据生成方法对比结果

为了验证 BSMOTE 方法的有效性，在不同扩充比例下将其与生成对抗网络（GAN，generative adversarial networks）做了对比。GAN 是一种采用对抗的思想来生成数据的方法，已经在图像生成等多个方面取得了成功应用。GAN 是由生成器和判别器构成。生成器负责生成与原始数据相似的数据，判别器负责判断该数据是生成数据还是真实数据。通过生成器和判别器的博弈，可以得到与原始数据相似的生成数据。

在该实验中，均衡数据集是由 BSMOTE 和 GAN 扩充得到的，且扩充前原始数据保持一致。并且考虑了不同扩充比例后即训练数据中生成数据与真实数据的比值不同的情况下的分类效果，实验结果如表 2 所示。

表 2　不同数据生成方法对比结果

扩充比例	方法	准确率/%	精确度/%	召回率/%	F1 分数
11:1	GAN	92.82	92.76	92.83	92.79
	BSMOTE	99.07	99.09	99.07	99.08
5:1	GAN	92.36	92.51	92.36	92.43
	BSMOTE	98.27	98.27	98.27	98.27
3:1	GAN	93.57	93.81	93.57	93.69
	BSMOTE	97.97	97.98	97.97	97.98
2:1	GAN	92.89	93.01	92.89	92.94
	BSMOTE	98.04	98.06	98.04	98.05
1:1	GAN	92.24	92.44	92.24	92.34
	BSMOTE	97.55	97.59	97.55	97.57

从表 2 中可以看出，BSMOTE 生成数据训练的模型检测效果优于 GAN。当扩充比例为 11∶1 时，BSMOTE 与 GAN 相比 4 个指标平均提高了 6.28%；当扩充比例为 5∶1 时，BSMOTE 与 GAN 相比 4 个指标平均提高了 5.86%；当扩充比例为 3∶1 时，BSMOTE 与 GAN 相比 4 个指标平均提高了 4.32%；当扩充比例为 2∶1 时，BSMOTE 与 GAN 相比 4 个指标平均提高了 5.12%；当扩充比例为 1∶1 时，BSMOTE 与 GAN 相比 4 个指标平均提高了 5.25%。

3.3.3　验证 GRU 的有效性

为了验证 GRU 对于用电数据分类的有效性，将其与 SVM 和 LSTM 做了对比。在该实验中，3 个模型所使用的数据集是 BSMOTE 扩充后的均衡数据集。实验中训练与测试数据集均一致，验证在该条件下不同方法的异常用电检测性能。并且在该实验中，还考虑了不同扩充比例时的分类效果，实验结果如表 3 所示。

从表 3 中可以得出，本文提出的方法的结果优于其他方法。当扩充比例为 1∶11 时，GRU 与 LSTM 相比 4 个评价指标提高了 3.40%～3.52%，与 SVM 相比提高了 1.52%～3.46%。当扩充比例为 1∶5 时，GRU 与 LSTM 相比 4 个评价指标提高了 5.4%～5.52%，与 SVM 相比提高了 3.00%～6.49%。当扩充比例为 1∶3 时，GRU 与 LSTM 相比 4 个评价指标提高了 4.69%～5.20%，与 SVM 相比提高了 2.82%～6.46%。当扩充比例为 1∶2 时，GRU 与 LSTM 相比 4 个评价指标提高了 5.64%～5.85%，与 SVM 相比提高了 3.00%～6.77%。当扩充比例为 1∶1 时，GRU 与 LSTM 相比 4 个评价指标提高了 5.60%～5.67%，与 SVM 相比提高了 3.16%～7.54%。

表 3　不同分类方法检测结果

扩充比例	方法	准确率/%	精确度/%	召回率/%	F1 分数
11:1	LSTM	95.54	95.69	95.55	95.62
	SVM	97.55	95.63	97.55	96.58
	GRU	99.07	99.09	99.07	99.08
5:1	LSTM	92.75	92.87	92.75	92.81
	SVM	95.27	91.78	95.27	93.49
	GRU	98.27	98.27	98.27	98.27
3:1	LSTM	92.78	93.29	92.77	92.99
	SVM	95.15	91.52	95.15	93.31
	GRU	97.97	97.98	97.97	97.98
2:1	LSTM	92.19	92.42	92.19	92.30
	SVM	95.04	91.29	95.04	93.13
	GRU	98.04	98.06	98.04	98.05
1:1	LSTM	91.88	91.99	91.89	91.94
	SVM	94.39	90.05	94.31	92.13
	GRU	97.55	97.59	97.55	97.57

4　结束语

本文提出了基于门控循环单元的非均衡数据驱动的异常用电检测方法。使用边界合成少数类过采样技术解决实际应用中异常用电数据过少导致的非均衡数据问题。边界合成过采样技术在数据类别边界生成数据，能够实现对少数类数据的有效扩充并且能够使得模型更容易学习不同类别数据的特征。为了更好地捕获用电数据的时间序列特征，采用 GRU 实现对用电数据的分类。经过详细的实验验证，表明该方法能够实现不同扩充比例情况下地数据有效扩充，并且能够以更高的准确率检测异常用电行为。在未来的研究中，将会致力于研究如何在保证检测准确率的基础上，简化模型，降低模型参数，并且进一步减少模型的训练时间。

参考文献：

[1] 苏才普．智能电网研究现状［J］．电气开关，2015，53（1）：1－4.

[2] MCDANIEL P，MCLAUGHLIN S. Security and privacy challenges in the smart grid［J］. IEEE Security and Privacy，2009，7（3）：75－77.

[3] 曹　峥．反窃电系统的研究与应用［D］．上海：上海交通大学，2011.

[4] 王全兴，李思韬．基于采集系统的反窃电技术分析及防范措施［J］．电测与仪表，2016，53（7）：78－83.

[5] 李端超，王　松，黄太贵，等．基于大数据平台的电网线损与窃电预警分析关键技术 [J]．电力系统保护与控制，2018，46 (5)：143－151.

[6] JúNIOR L A P，RAMOS C C O，RODRIGUES D，et al. Unsupervised non-technical losses identification through optimum-path forest [J]．Electric Power Systems Research，2016，140：413－423.

[7] BABU T V，MURTHY T S，SIVAIAH B. Detecting unusual customer consumption profiles in power distribution systems—APSPDCL [C] //2013 IEEE International Conference on Computational Intelligence and Computing Research，IEEE，2013：1－5.

[8] DE NADAI M，VAN SOMEREN M. Short－term anomaly detection in gas consumption through ARIMA and artificial neural network forecast [C] //2015 IEEE Workshop on Environmental，Energy，and Structural Monitoring Systems (EESMS) Proceedings. IEEE，2015：250－255.

[9] CHOU J S，TELAGA A S. Real-time detection of anomalous power consumption [J]．Renewable and Sustainable Energy Reviews，2014，33：400－411.

[10] SIAL A，SINGH A，MAHANTI A. Detecting anomalous energy consumption using contextual analysis of smart meter data [J]．Wireless Networks，2021，27 (6)：4275－4292.

[11] JAKKULA V，COOK D. Outlier detection in smart environment structured power datasets [C] //2010 Sixth International Conference on Intelligent Environments，IEEE，2010：29－33.

[12] DEPURU S S S R，WANG L，DEVABHAKTUNI V. Support vector machine based data classification for detection of electricity theft [C] //2011 IEEE/PES Power Systems Conference and Exposition，IEEE，2011：1－8.

[13] YIP S C，TAN W N，TAN C K，et al. An anomaly detection framework for identifying energy theft and defective meters in smart grids [J]．International Journal of Electrical Power & Energy Systems，2018，101：189－203.

[14] REIF M，GOLDSTEIN M，STAHL A，et al. Anomaly detection by combining decision trees and parametric densities [C] //2008 19th International Conference on Pattern Recognition，IEEE，2008：1－4.

[15] KIM T，LEE D，CHOI J，et al. Extracting baseline electricity usage using gradient tree boosting [C] //2015 IEEE International Conference on Smart City/Socialcom/Sustaincom (SmartCity)，IEEE，2015：734－741.

[16] 许　刚，谈元鹏，戴腾辉．稀疏随机森林下的用电侧异常行为模式检测 [J]．电网技术，2017，41 (6)：1964－1973.

[17] HOLLINGSWORTH K，ROUSE K，CHO J，et al. Energy anomaly detection with forecasting and deep learning [C] //2018 IEEE International Conference on Big Data (Big Data)，IEEE，2018：4921－4925.

[18] CHAHLA C，SNOUSSI H，MERGHEM L，et al. A novel approach for anomaly detection in power consumption data [C] //ICPRAM，2019：483－490.

[19] WENG Y，ZHANG N，XIA C. Multi-agent-based unsupervised detection of energy consumption anomalies on smart campus [J]．IEEE Access，2018，7：2169－2178.

[20] 林女贵，洪兰秀，黄道姗，等．基于改进深度自编码网络的异常用电行为辨识 [J]．中国电力，2020，53 (6)：18－26.

[21] PEREIRA J，SILVEIRA M. Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention [C] //2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)，IEEE，2018：1275－1282.

[22] LI S，HAN Y，YAO X，et al. Electricity theft detection in power grids with deep learning and random forests [J]．Journal of Electrical and Computer Engineering，2019：1－12.

[23] ZHENG Z，YANG Y，NIU X，et al. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids [J]．IEEE Transactions on Industrial Informatics，2017，14 (4)：1606－1615.

[24] TANG Z，CHEN Z，BAO Y，et al. Convolutional neural network－based data anomaly detection method using multiple information for structural health monitoring [J]．Structural Control and Health Monitoring，2019，26 (1)：e2296.

[25] LI K，REN B，GUAN T，et al. A hybrid cluster-borderline SMOTE method for imbalanced data of rock groutability classification [J]．Bulletin of Engineering Geology and the Environment，2022，81 (1)：1－15.

[26] SHEN C，ZHANG H，MENG S，et al. Augmented data driven self-attention deep learning method for imbalanced fault diagnosis of the HVAC chiller [J]．Engineering Applications of Artificial Intelligence，2023，117：105540.

[27] KHODABANDELOU G，JUNG P G，AMIRAT Y，et al. Attention-based gated recurrent unit for gesture recognition [J]．IEEE Transactions on Automation Science and Engineering，2020，18 (2)：495－507.

[28] FENG R，JIANG W，YU N，et al. Projected minimal gated recurrent unit for speech recognition [J]．IEEE Access，2020，8：215192－215201.

[29] MENG S，LI C，PENG W，et al. Empirical mode decomposition-based multi-scale spectral graph convolution network for abnormal electricity consumption detection [J]．Neural Computing and Applications，2023：1－17.

[30] 黄宇斐，石新发，贺石中，等．一种基于主成分分析与支持向量机的风电齿轮箱故障诊断方法 [J]．热能动力工程，2022，37 (10)：175－181.

[31] 余畅文，潘万宝，刘　练，等．基于斑点鬣狗算法优化支持向量机的短期风电功率预测 [J]．电工技术，2022 (15)：4－6.

[32] 李静茹，姚　方．引入注意力机制的 CNN 和 LSTM 复合风电预测模型 [J]．电气自动化，2022，44 (6)：4－6.