研究与开发

基于 GAN 数据重构的电信用户流失预测方法

阿克弘，胡晓东

（中国电信股份有限公司西宁分公司，青海 西宁 810001）

摘　要：用户是运营商利益的核心。随着携号转网政策的出台，运营商之间的竞争越发激烈。为了提前精准有效地预测用户流失倾向，提出了一种基于生成对抗网络（generative adversarial network，GAN）数据重构的电信用户流失预测方法。首先，利用有效的数据预处理方法电信用户流失数据中的脏数据；其次，利用 GAN 重构电信用户流失数据，解决电信用户流失数据不平衡问题；最后，利用极度梯度提升树（extreme gradient boosting，XGBoost）算法分别训练基于 GAN 重构的电信用户流失预测模型和基于合成少数类过采样技术（synthetic minority oversampling technique，SMOTE）采样的电信用户流失预测模型，对比两种模型的预测精度。实验结果表明，GAN 重构后的电信用户流失预测模型预测精度比未重构的预测模型的准确率提升了 6.75%，查准率提升了 25.91%，召回率提升了 30.91%，*F*1 值提升了 28.73%。该方法能够有效提升电信用户流失预测的准确度。

关键词：XGBoost 算法；生成对抗网络；用户流失；数据重构；SMOTE

中图分类号：TN91, TP181

文献标志码：A

doi: 10.11959/j.issn.1000–0801.2023038

GAN data reconstruction based prediction method of telecom subscriber loss

A Kehong, HU Xiaodong

Xining Branch of China Telecom Co., Ltd., Xining 810001, China

Abstract: Users are the core of operators' interests. With the introduction of the policy of transferring network with a number, the competition between operators becomes more and more fierce. In order to accurately predict subscriber loss tendency in advance, a prediction method of subscriber loss based on generative adversarial network data reconstruction was proposed. Firstly, the dirty data in the telecom subscriber loss data was used by effective data preprocessing method. Secondly, the GAN was used to reconstruct the telecom subscriber loss data to solve the problem of the imbalance of the telecom subscriber loss data. Finally, extreme gradient boosting algorithm was used to train the telecom subscriber loss prediction model based on GAN reconstruction and the SMOTE sampling model based on synthetic minority oversampling technique sampling method respectively, and compare the prediction accuracy of the two models. The experimental results show that the prediction accuracy of the GAN reconstructed telecom subscriber loss prediction model is increased by 6.75%, the accuracy rate is increased by 25.91%, the recall rate is increased by

收稿日期：2022–12–17；修回日期：2023–03–01

30.91%, and the *F*1-score is increased by 28.73% compared with the unreconstructed prediction model. This method can effectively improve the accuracy of telecom subscriber loss prediction.

Key words: XGBoost algorithm, generative adversarial network, customer churn, data reconstruction, synthetic minority oversampling technique

0 引言

随着信息化进入以大数据、云计算、5G、物联网为表征的新阶段，全球正经历着一场新的深度信息化浪潮。在深度信息化时代，全球电信用户数量持续高速增长，电信行业竞争越发激烈，用户作为运营商的利益来源的核心点，谁拥有的用户多，谁的市场占有率就高。随着“携号转网”“降费提速”等一系列政策的出台，运营商之间竞争愈发激烈，如何减少用户的流失对于运营商而言变得尤为重要。

目前，对电信用户流失的研究主要基于人工智能算法，将人工智能算法运用到电信行业的用户分析中，利用数据挖掘方法挖掘其中潜藏的高价值信息并加以正确利用，可以有效降低用户流失度，对于运营商具有重要的理论意义和实践价值。文献[1]提出了分群构建电信用户流失预警模型，通过皮尔逊（Pearson）相关性分析筛选出与电信用户相关性强的特征，采用 *K*-means 聚类算法进行分群，采用极度梯度提升树（XGBoost）等聚类算法对分群预测用户流失概率，并比较各算法的优劣。文献[2]提出了基于图算法的自动特征交叉构造的框架，用来挖掘交叉特征，后训练预测模型做准备，基于时序图注意力机制的在线算法实现在线预测。文献[3]利用 Fisher 判别方程和逻辑回归建立电信用户流失预测模型，具有较高的准确率。文献[4]根据业务经验，由用户流失原因反推导致用户流失的因子，根据推导的因子和深度学习算法、卷积神经网络（CNN）预测用户流失概率。文献[5]利用逻辑回归和随机森林模型筛选与用户流失相关性强的特征，在利用机器学习算法训练预测模型。

以上研究在不同程度上对电信用户流失度预测起到了推动作用，为预测电信用户流失提供了新的思路和方法，但普遍在电信用户数据不平衡的问题上没有深入研究。本文基于 Kaggle 的电信用户流失数据集，在目前的研究成果基础上提出了基于生成对抗网络（generative adversarial network，GAN）数据重构的电信用户流失预测方法，为电信用户流失数据样本不平衡的问题提供新的解决思路，提升电信用户流失预测模型的准确率。

1 建模思路和方法

电信用户流失预测的本质是利用人工智能算法和数据挖掘技术对以往电信用户的个人信息和业务信息进行分析，如套餐费用、消费记录、是否流失等相关数据进行学习，训练出分类预测模型。然后利用训练好的模型预测用户流失的概率。对流失概率高的用户及时采取相应的挽留措施，降低用户离网率。

（1）对基于 Kaggle 的电信用户流失数据集进行预处理。由于系统宕机、数据传输错误等因素，用户流失数据中夹杂着很多脏数据。因此，首先去除用户流失数据中的重复值、缺失值以提升后续建立模型的准确性；然后分箱离散化，对非数字类型特征进行哑变量处理，删除异常值；最后筛选与电信用户流失相关性强的特征，并统一量纲做归一化处理。

（2）利用 GAN 生成与电信用户流失数据分布相似的数据集，重构电信用户流失数据集，用来解决电信用户流失数据中用户流失数据和未流失用户数据样本不平衡的问题。

（3）分别训练基于门控循环单元（GRU）生成对抗网络的电信用户流失预测模型，并对比分

析重构和未重构数据集所训练电信用户预测模型的准确度。采用常用来解决数据不平衡的 SMOTE 方法采样电信用户流失数据集训练电信用户流失预测模型，对比两种解决数据样本不平衡方法的优劣。

2　数据准备

2.1　数据说明

本文采用的数据是基于 Kaggle 的电信用户流失数据集。电信用户流失数据集中记录了用户个人信息、账户信息、服务信息以及最为核心是否流失的标签数据等 21 个有效指标。收集的电信用户流失数据集见表 1。

表 1　收集的电信用户流失数据集

性别	配偶	…	总费用/美元	是否流失
女	有	…	29.85	是
女	无	…	1 889.5	否
…	…	…	…	…
男	有	…	306.6	是
男	无	…	6 844.5	否

2.2　数据预处理

由于系统宕机、人为误操作、网络故障等因素，电信用户流失数据中夹杂脏数据，如存在数据缺失值、重复值、异常值等。这些脏数据会影响预测模型的准确度，需要在训练模型和测试前进行数据预处理，过程如下。

（1）数据缺失值处理：缺失值只有 11 条，对模型整体影响较小，予以删除。

（2）数据重复值处理：重复值对模型精度无益反而有害，予以删除。

（3）数据哑变量处理：将非数值类型对象类转为数值类型。

（4）数据异常值处理：利用三西格玛准则对异常值予以删除，检测数据值落入区间 $[\mu-3\sigma,\mu+3\sigma]$ 之外时为极小概率事件，属于粗大误差数据应予以剔除。其中，μ为特征数据的均值，σ为特征数据的标准差。

（5）特征筛选：电信用户流失数据集中的每一个标签并非都与用户流失有关，为了提高预测模型的准确度和减少模型的训练时长，选取对数据分布没有特殊要求、可解释性好和相关系数收敛快的斯皮尔曼相关性分析方法提取与电信用户流失关联性强的特征子集。筛选的特征子集见表 2。

（6）数据归一化处理：筛选的数据中各个特征之间的量纲和量纲单位不同，为了消除特征之间的量纲影响和减少异常值的影响，需要进行数据归一化处理，将特征数据进行归一化到(–1,1)之间，计算式如下。

$$\overline{x}_{A_i}=\frac{x_{A_i}-\mathrm{avg}(A_i)}{\max(A_i)-\min(A_i)} \tag{1}$$

其中，x_{A_i} 为样本 x 在特征 A_i 上的数据值，$i=(1,2,3,\cdots,12)$；$\mathrm{avg}(A_i)$ 为特征 A_i 的均值，$\max(A_i)$ 为特征数据 A_i 的最大值，$\min(A_i)$ 为特征数据 A_i 的最小值，$\overline{x}_{A_i}$ 归一化以后的数据 x_{A_i}。

表 2　筛选的特征子集

选取特征名称	斯皮尔曼相关性系数	特征编号
Contract	–0.406 2	A_1
tenure	–0.367 0	A_2
OnlineSecurity	–0.303 9	A_3
TechSupport	–0.296 8	A_4
TotalCharges	–0.229 9	A_5
OnlineBackup	–0.203 2	A_6
DeviceProtection	–0.185 9	A_7
Dependents	–0.164 2	A_8
Partner	–0.155 9	A_9
SeniorCitizen	0.150 9	A_{10}
MonthlyCharges	0.184 7	A_{11}
PaperlessBilling	0.191 8	A_{12}

3 基于 GAN 重构数据

GAN 由生成器（generator）和判别器（discriminator）组成，是一种无监督学习算法。生成器负责生成模拟数据，并不断优化生成的数据，达到以假乱真的效果。而判别器负责判断输入的数据是真实数据还是生成数据，不断提高判别能力，两者形成对抗。

GAN 的目标函数为：

$$\min_G \max_D V(D,G) = E[\lg P(S=\text{real}|X_{\text{real}})] + E[\lg P(S=\text{fake}|X_{\text{fake}})] \quad (2)$$

其中，生成器目标函数为：

$$\min_G V(G) = E[\lg P(S=\text{fake}|X_{\text{fake}})] \quad (3)$$

判别器的目标函数为：

$$\max_D V(D) = E[\lg P(S=\text{real}|X_{\text{real}})] + E[\lg P(S=\text{fake}|X_{\text{fake}})] \quad (4)$$

向 GAN 的生成器输入随机数，生成器 G 生成模拟数据 $X_{\text{fake}}=G(z)$，将 X_{fake} 或真实样本数据 X_{real} 输入判别器 D 中判定输入样本是否真实的概率为 $P(S|X)=D(X)$。GAN 训练过程中，G 和 D 交替训练，相互博弈，直到达到纳什均衡[6]。GAN 的基本框架如图 1 所示。

GAN 的生成器和判别器均采用循环神经网络（recurrent neural network，RNN），输入层神经单元为 20 个，隐含层设置为 5 层，输出层神经单元为 20 个，激活函数为 sigmoid，均方误差作为损失函数。选择基于 Kaggle 的电信用户流失数据集，筛选出电信流失用户数据和未流失用户数据，未流失用户数据有 5 174 条，流失用户数据有 1 869 条，利用 GAN 生成与流失用户数据分布相似的数据 3 295 条，使得电信流失用户数据集与未流失用户数据集样本数量一致。

4 基于 XGBoost 算法训练预测模型

4.1 XGBoost 算法

XGBoost 算法是在自适应增强（adaptive boosting，adaBoost）算法和梯度提升迭代决策树（gradient boosting decision tree，GBDT）算法基础上优化形成的，能有效地处理分类和回归问题。XGBoost 算法由陈天齐设计，致力于让提升树突破自身的计算极限，以实现运算快速、性能优越的工程目标[7]。和传统的梯度提升算法相比，XGBoost 算法更加快速，并且已经被认为是在分类和回归上都拥有超高性能的先进评估器[8-9]。XGBoost 的目标函数为传统损失函数加模型复杂度[10]。

$$L(\phi) = \sum_{i=1}^{m} l(y_i, \overline{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (5)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\|\omega\|^2 \quad (6)$$

其中，$\sum_{i=1}^{\mathrm{m}} l(\mathrm{y}_i, \overline{y}_i)$ 为用户流失预测模型的训练误差，i 为正则化项，i 代表用户流失数据集中的 i 个

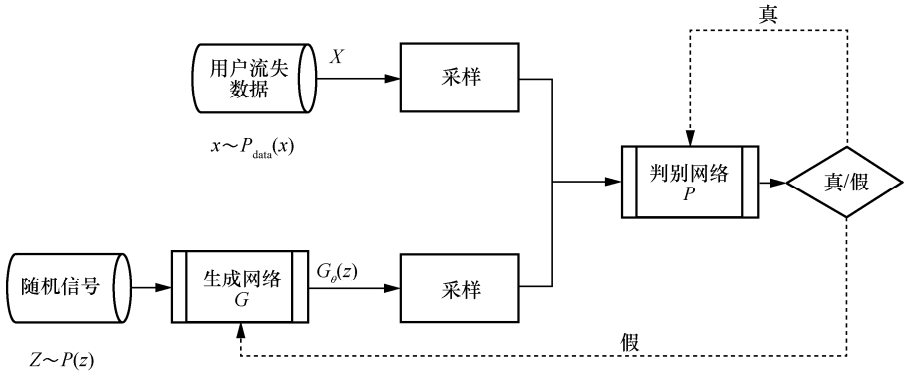

图 1 GAN 的基本框架

样本，m 代表导入第 k 棵树的数据总量，K 代表建立所有树。在迭代每棵树的时候，都最小化 $L(\phi)$ 来获得最优的 $\overline{y_i}$，同时最小化了模型的错误率和模型的复杂度；T 为树叶子节点数，ω 为叶子权重值，γ 为叶子数惩罚正则项，λ 为叶子权重惩罚正则项。每次建树优化如下目标：

$$\overline{L}^{(t)}=\sum_{i=1}^{n}[g_i f_t(x_i)+\frac{1}{2}h_t f_t^2(x_i)]+\Omega(f_t) \tag{7}$$

$$g_i=\partial_{\overline{y}_i} l\left(y_i,\overline{y}_i^{(t-1)}\right) \tag{8}$$

$$h_i=\partial^2_{\overline{y}_i^{(t-1)}} l\left(y_i,\overline{y}_i^{(t-1)}\right) \tag{9}$$

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^{\ 2} \tag{10}$$

最终得到叶子权重值为：

$$\omega_j^{\ *}=-\frac{G_j}{H_j+\lambda} \tag{11}$$

其中，G_j 叶子节点 j 所包含样本的一阶偏导数累加之和，H_j 叶子节点 j 所包含样本的二阶偏导数累加之和。最终的目标值为：

$$\overline{L}^{*}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^{\ 2}}{H_j+\lambda}+\gamma T \tag{12}$$

4.2　实验环境的设置

本文所有实验运行环境均为：操作系统为 Windows10、Python 版本为 3.8.13、集成开发环境为 Anaconda3。本文所用到的各种算法调用的是开源机器学习框架 TensorFlow 和 Sklearn。

4.3　训练电信用户流预测模型

将没有经过 GAN 数据重构的基于 Kaggle 的电信用户流失数据集进行数据预处理，选择数据的 70%作为电信用户流失预测模型的训练集，30%作为测试集，模型的评估指标为准确率（accuracy）、查准率（precision）、召回率（recall）和 $F1$ 值（$F1$-score）。准确率为分类正确的样本总数占样本总数的比例，其计算式为：

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{13}$$

查准率为分类正确的正样本数（TP）占预测为正样本的总样本数（TP+FP）的比例，其计算式为：

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{14}$$

召回率为分类正确的正样本数（TP）占实际的正样本总数（TP+FN）的比例，其计算式为：

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{15}$$

$F1$ 值为查准率和召回率的加权调和平均，其计算式为：

$$F1\text{-score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{16}$$

其中，TP 表示真正，即把正样本（P）正确地分类为正类（P）；TN 表示真负，即把负样本（N）正确地分类为正类（T）；FN 表示假负，即把正样本（P）错误地分类为正类（T）；FP 表示假正，即把负样本（N）错误地分类为正类（P）。

训练电信用户流预测模型中采用 XGBoost、随机森林、邻近算法（*K*-nearest neighbor，*K*NN)、逻辑回归、支持向量机（support vector machine，SVM）等算法进行训练，并通过十折交叉验证结果对比分析效果最优的算法。经过训练，基于各种算法的电信用户流失预测模型在测试集上的评价效果见表 3。

表 3　基于各种算法的电信用户流失预测模型在测试集上的评价效果

算法	准确率	查准率	召回率	$F1$ 值
XGBoost	0.793 2	0.622 4	0.533 4	0.574 5
随机森林	0.787 0	0.621 1	0.477 4	0.539 8
*K*NN	0.776 6	0.587 0	0.493 6	0.536 3
逻辑回归	0.791 2	0.613 6	0.528 0	0.570 7
支持向量机	0.789 2	0.631 4	0.459 3	0.535 3

通过以上算法训练电信用户流失预测模型后，在测试集上的十折交叉验证表现结果可看出 XGBoost 的查准率比支持向量机低 0.01 左右，其

他的各项评价指标均高于其他算法的评价指标。其中，基于 XGBoost 算法的电信用户流失预测模型在测试集上的混淆矩阵如图 2 所示，混淆矩阵 1 代表预测为真，0 代表预测为假。在测试集上的准确度为 79.32%。可见 XGBoost 算法在电信用户流失预测模型上具有较高的准确度。

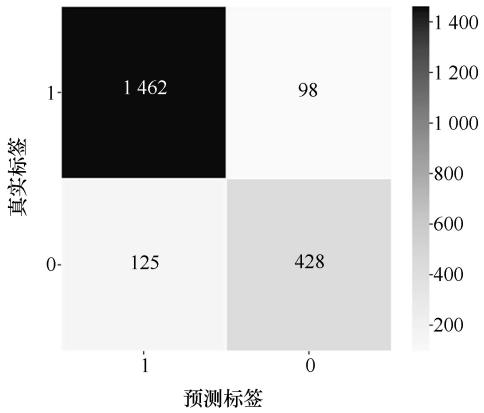

图 2　基于 XGBoost 算法的电信用户流失预测模型在测试集上的混淆矩阵

为了解决电信用户数据不平衡问题，很多学者利用 SMOTE 合成少数类过采样技术，该算法是在随机采样的基础上改进的一种过采样算法[11-13]。采样过程如下。

步骤 1　找到少数类中的某个样本 X_i 的 K 个近邻。

步骤 2　随机选取一个近邻 $X_i(n,n)$，并随机生成一个介于 0～1 的数，合成新的样本。

步骤 3　对每个随机选出的 X_i 的近邻重复步骤 2。

步骤 4　对少数类的每一个样本重复上述步骤。

5　实例应用验证

首先将上述 GAN 重构的数据按 7:3 比例分组，70%作为电信用户流失预测模型的训练集，30%作为测试集，利用 XGBoost 算法和经过 GAN 数据重构的电信用户流失数据集训练电信用户流失预测模型，测试集通过十折交叉验证算法的各评价指标见表 4。

表 4　测试集通过十折交叉验证算法的各评价指标

算法	准确率	查准率	召回率	*F*1 值
XGBoost	0.793 2	0.622 4	0.533 4	0.574 5
SMOTE-XGBoost	0.824 0	0.689 9	0.558 6	0.587 3
GAN-XGBoost	0.860 7	0.835 3	0.755 9	0.793 75

可见，经过 GAN 数据重构后的电信用户流失预测模型的预测性能有了很大的提升，基于 GAN-XGBoost 算法的电信用户流失预测模型在测试集上的混淆矩阵如图 3 所示，在测试集上的准确率提升了 6.75%，查准率提升了 25.91%，召回率提升了 30.91%，*F*1 值提升了 28.73%。说明该方法对提升电信用户流失预测准确率有很大的作用。

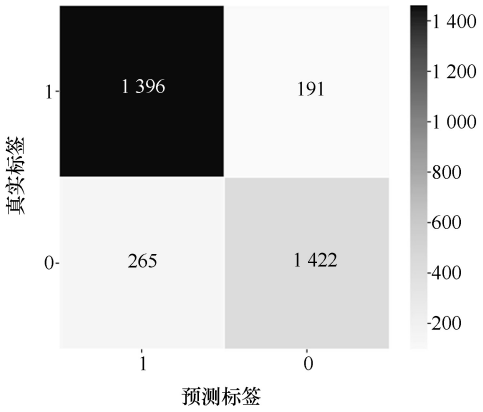

图 3　基于 GAN-XGBoost 算法的电信用户流失预测模型在测试集上的混淆矩阵

接下来，对比分析利用 GAN 和利用 SMOTE 过采样两种解决数据样本不平衡问题方法的优劣。两种方法采用相同的数据集做实验，首先对原始数据进行数据预处理，然后采用 SMOTE 过采样方法进行采样，利用 XGBoost 算法和经过 SMOTE 过采样的电信用户流失数据集训练电信用户流失预测模型，数据还是按 7:3 划分，70%作为电信用户流失预测模型训练集，30%作为测试集。模型训练结束后，利用十折交叉验证方法测试模型在测试集上表现，实验结果见表 4。基于 SMOTE 过采样和 XGBoost 算法训练的电信用户流失预测模型在测试集在准确率提

升了 3.08%，查准率提升了 6.75%，召回率提升了 2.52%，*F*1 值提升了 1.28%。这说明该方法对提升预测准确率也有一定的作用。但是基于 SMOTE-XGBoost 的电信用户流失预测模型的效果略低于 GAN-XGBoost 的电信用户流失预测模型。可见经过 GAN 数据重构后的电信用户流失预测模型的预测性能有了很大的提升。

6　结束语

本文利用基于 Kaggle 的电信用户流失数据集和 XGBoost 算法训练电信用户流失预测模型。针对电信用户流失数据集中流失样本数据和未流失样本数据不均衡的问题，利用 GAN 生成与流失数据集分布一致的数据，重构电信用户流失数据集。然后利用重构的数据样本和 XGBoost 算法训练电信用户流失预测模型。与此同时，利用 SMOTE 过采样方法和 XGBoost 算法训练电信用户流失预测模型，对比分析两种方法的优劣。

（1）训练电信用户流预测模型中采用 XGBoost、随机森林、*K*NN、逻辑回归、支持向量机等算法进行训练，并通过十折交叉验证结果对比分析几种算法的表现，得到基于 XGBoost 算法训练的电信用户流失预测模型效果最好的结论。

（2）利用 GAN 生成电信用户数据集中的与流失用户数据分布相似的数据，重构电信用户数据集，使数据集中流失用户数据样本和未流失数据样本的数据量一致。利用重构后的数据集和 XGBoost 算法训练电信用户流失预测模型。与未重构数据训练的电信用户流失预测模型相比，重构后的模型在测试集上的准确率提升了 6.75%，查准率提升了 25.91%，召回率提升了 30.91%，*F*1 值提升了 28.73%。

（3）为了验证 GAN 方法的优劣，利用 SMOTE 过采样方法和 XGBoost 算法训练电信用户流失预测模型。训练后的预测模型在测试集上的准确率提升了 3.08%，查准率提升了 6.75%，召回率提升了 2.52%，*F*1 值提升了 1.28%。该方法虽然对提升电信用户流失预测模型也有一定作用，但是效果不如 GAN 数据重构法。

参考文献：

[1] 叶艳艳. 基于机器学习的电信用户流失预警模型预测与分析[D]. 济南: 山东大学, 2021.
YE Y Y. Prediction and analysis of telecom customer churn warning model based on machine learning[D]. Jinan: Shandong University, 2021.

[2] 钟保权. 基于图学习的电信用户流失预测算法[D]. 杭州: 浙江大学, 2021.
ZHONG B Q. Graph learning algorithm for telecom user churn prediction[D]. Hangzhou: Zhejiang University, 2021.

[3] 乔健, 诸佳慧, 严康桓. 基于随机森林 CART 特征选择改进算法的电信客户流失预测模型[J]. 电信工程技术与标准化, 2022, 35(3): 78-82.
QIAO J, ZHU J H, YAN K H. Telecom customer churn prediction model based on improved random forest cart feature selection algorithm[J]. Telecom Engineering Technics and Standardization, 2022, 35(3): 78-82.

[4] ZHANG T Y, MORO S, RAMOS R F. A data-driven approach to improve customer churn prediction based on telecom customer segmentation[J]. Future Internet, 2022, 14(3): 94.

[5] JAIN H, KHUNTETA A, SRIVASTAVA S. Telecom churn prediction and used techniques, datasets and performance measures: a review[J]. Telecommunication Systems, 2021, 76(4): 613-630.

[6] 郑声晟, 殷海兵, 黄晓峰, 等. 基于 GAN 的无监督域自适应行人重识别[J]. 电信科学, 2021, 37(2): 99-106.
ZHENG S S, YIN H B, HUANG X F, et al. GAN-based unsupervised domain adaptive person re-identification[J]. Telecommunication Science, 2021, 37(2): 99-106.

[7] ERIA K, MARIKANNAN B P. Significance-based feature extraction for customer churn prediction data in the telecom sector[J]. Journal of Computational and Theoretical Nanoscience, 2019, 16(8): 3428-3431.

[8] LIU H Z, ZHANG X, SHEN X W, et al. A fair and efficient hybrid federated learning framework based on XGBoost for distributed power prediction[J]. arXiv preprint, 2022, arXiv: 2201.02783.

[9] 王愈轩, 梁沁雯, 章思远, 等. 基于 LSTM-XGBoost 组合的超短期风电功率预测方法[J]. 科学技术与工程, 2022, 22(14): 5629-5635.
WANG Y X, LIANG Q W, ZHANG S Y, et al. An ultra-short-term wind power prediction method based on LSTM-XGBoost combination[J]. Science Technology and Engineering, 2022, 22(14): 5629-5635.

[10] LIN M Y, ZHU X F, HUA T, et al. Detection of ionospheric scintillation based on XGBoost model improved by SMOTE-ENN technique[J]. Remote Sensing, 2021, 13(13): 2577.

[11] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[M]//Lecture Notes in Computer Science. Heidelberg: Springer, 2005: 878-887.

[12] 张晨路. 基于 G-SMOTE 和 Biased-SVM 的内部威胁用户检测[J]. 中北大学学报(自然科学版), 2022, 43(2): 147-152.
ZHANG C L. User detection of insider threat detection based on G-SMOTE and biased-SVM[J]. Journal of North University of China (Natural Science Edition), 2022, 43(2): 147-152.

[13] 石洪波, 陈雨文, 陈鑫. SMOTE 过采样及其改进算法研究综述[J]. 智能系统学报, 2019, 14(6): 1073-1083.
SHI H B, CHEN Y W, CHEN X. Summary of research on SMOTE oversampling and its improved algorithms[J]. CAAI Transactions on Intelligent Systems, 2019, 14(6): 1073-1083.

[作者简介]

阿克弘（1991– ），男，中国电信股份有限公司西宁分公司工程师、产品部主任，主要研究方向为基于电信用户数据的数据分析及数据挖掘。

胡晓东（1995– ），男，中国电信股份有限公司西宁分公司助理工程师，主要研究方向为基于风电机组运行数据的数据挖掘及故障预警、基于电信用户数据的数据分析及数据挖掘。