doi:10.3772/j.issn.2095-915x.2018.03.014

半监督学习的用电客户价值评价建模

刘刚[1] 高迪[1] 付薇薇[2] 刘霞[3] 刘欣[2]

1. 国网冀北电力有限公司运营监测（控）中心 北京 100056；
2. 北京科技大学计算机与通信工程学院 北京 100083；
3. 北京博望华科科技有限公司 北京 100045

摘 要 随着中国售电市场的逐渐开放，大型垄断型电力企业也要加入到激烈的市场竞争中，对用电客户的价值评价逐渐成为电力企业工作的重点。本文结合半监督的机器学习算法，提出基于电力业务行为预测的用电客户价值评价模型，将价值评价问题转化为分类问题。借助随机森林和半监督随机森林算法构建预测模型，对与价值强相关的电力业务行为进行准确预测，降低类不平衡问题对预测模型的影响，为用电客户价值评价提供了一个新思路。

关键词： 价值评价；电力业务；行为预测；对抗生成网络；半监督随机森林

中图分类号： TP311

开放科学（资源服务）标识码（OSID）

Electricity Customer Value Evaluation Modeling Based on Semi-supervised Learning

LIU Gang[1] GAO Di[1] FU Weiwei[2] LIU Xia[3] LIU Xin[2]

1. Operation Monitor Center, State Grid Jibei Electric Power Company, Beijing 100056, China;
2. School of Computer and Communication Engineering, University of Science & Technology Beijing, Beijing 100083, China;
3. Beijing Bowang Huake Technology Co. Ltd., Beijing 100045, China

基金项目： 国家科技支撑计划项目（2017YFB1002304）。

作者简介： 刘刚（1977–），硕士，高级工程师，研究方向：大数据分析应用及预测预警模型、跨专业间运营协同效率提升、电网运营稽查方法；高迪（1985–），硕士，高级工程师，研究方向：大数据分析、预测预警建模、运营协同效率优化；付薇薇（1992–），硕士生，研究方向：人工智能、大数据处理；刘霞（1990–），学士，工程师，研究方向：电力行业信息系统实施、运行、数据库管理、系统的运行维护、数据分析；刘欣（1987–），博士，讲师，研究方向：智能信息处理、挖掘与决策，E–mail：liuxin@ustb.edu.cn。

doi:10.3772/j.issn.2095-915x.2018.03.014

Abstract With the gradual opening up of Chinese electricity market, large-scale monopoly power enterprises are forced to join in the fierce competition. The evaluation of electricity users becomes more and more important. This paper proposes an evaluation model based on electric business behaviors using semi-supervised machine learning algorithms, and changes the evaluation problems to classification problems. By building predict models using random forest and semi-random forest algorithm, it can decrease the influence of class imbalance problems and give an accurate prediction of user business behaviors, which may provide a new method to evaluate the electricity user value.

Keywords: Value evaluation; electric power; behavior prediction; generative adversarial network; semi-supervised random forest

1 引言

用电客户的价值评价是当前电力企业工作的关注重点[1-2]，相关专家学者对用电客户的评价模型研究主要集中在：1）基于不同用电客户价值定义的价值评价指标体系研究[3]；2）基于海量用电客户数据的用电客户价值挖掘。近期机器学习的广泛应用，推动了基于海量数据的用电客户价值评价建模的研究与发展[4-5]。机器学习在价值评价建模中的研究包括：1）基于聚类算法的用电客户价值划分。此类研究通过K均值（K-means）及其改进的聚类算法对用电客户的月用电量、欠费次数、违约用电次数、功率系数、峰谷用电占比等指标进行聚类划分，通过对聚类结果的业务分析得到用电客户的价值评价[6-9]。2）基于分类算法的用电客户价值预测。此类研究采用专家评分法、主成分分析法、层次分析法(Analytic Hierarchy Process，AHP)等不同的价值评价指标体系，对部分用电客户进行价值区分，并以之为目标变量，采用有监督的分类算法构建分类模型，对剩余的用电客户的价值进行分类[3,10-13]。

然而，目前用电客户价值评价建模中存在针对高维度特征指标赋权误差大，随着数据维度及层级的增加，模型构建和扩展难度随之增大的问题，且考虑潜在高价值用电客户发生欠费、违约行为可能性的评价模型较少。针对上述问题，本文提出一种基于半监督学习的用电客户价值评价模型，该模型将价值评价问题转化为分类问题，通过机器学习算法对与价值强相关的电力业务行为进行预测，较好地避免了多层级、多指标的评价体系赋权问题，进而提高了模型的客观性。此外，根据用电客户的业务行为来计算价值评分，可更加直接地反映出用电客户主体与客户价值间的关联关系。

2 基于电力业务行为的用电客户价值评价模型

电力营销数据可以分为两类重要数据，静态属性数据和动态行为数据。静态属性数据为

电力用户较为稳定的信息，如，电压等级、用电规模、行业等，一般相对稳定。动态行为数据，即，用户不断变化的行为信息，例如，增容容量大小、违约行为类型、缴费金额等。静态属性数据主要描述用电客户的属性特征，而动态行为数据则描述了用电客户的各种电力业务行为。经前期调研发现，电力行业数据可被抽象为多个业务行为，每个业务行为都是高内聚、低耦合的特征指标。

本文根据冀北电力公司真实的营销数据，将其抽象为变损用电、线损用电、违约用电、投诉、用电检查、窃电等业务行为。这些业务行为按照是否定期发生可分为“定期业务行为”(如：“普通用电行为”、“缴费行为”等几乎每个月都会定期发生的行为)和“非定期业务行为”(如：“欠费行为”、“预付行为”等)，如图1所示。传统的用电客户价值评价模型使用“发展潜力”、“信用度”、“忠诚度”等来评价用电客户的潜在价值，预测用电客户未来违约、欠费、由于经营不善用电量下降的可能性，实际上就是根据用电客户具体的电力业务行为来评价其价值。

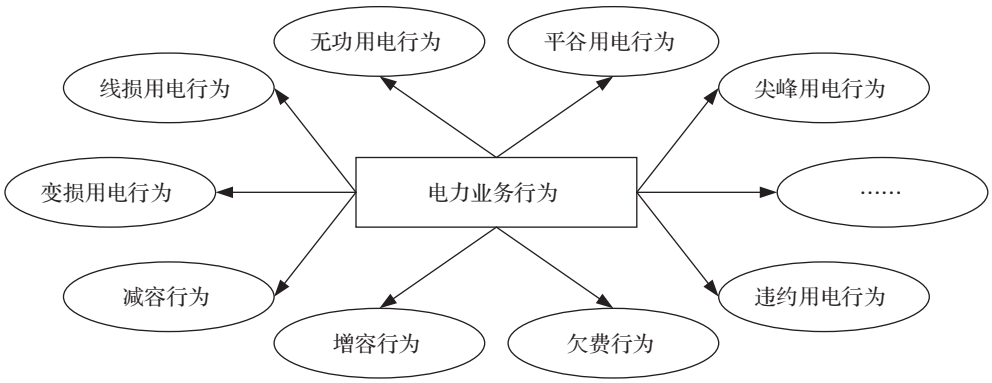

图 1　电力业务行为

表 1　电力业务行为预测模型的目标变量及其划分方式

模型	模型目标定义	划分方式
预付能力	半年内用电客户预付金额占比	聚类
用电需求	用电客户月平均用电量	聚类
欠费风险	半年内用电客户的欠费金额	聚类
安全风险	半年内用电客户出现安全隐患的次数	“是”或“否”
欺诈风险	半年内用电客户出现违约用电的次数	“是”或“否”
增容趋势	半年内用电客户是否会增容	“是”或“否”
减容趋势	半年内用电客户是否会减容	“是”或“否”
销户风险	半年内用电客户是否会销户	“是”或“否”

本文提出一种基于电力业务行为的构建价值评价指标体系，通过构建机器学习预测模型来预测用电客户的预付能力、用电需求、欠费风险、欺诈风险、增容趋势、减容趋势、销户风险。以电力营销数据的统计分析为依据，结合电力通用业务，将预测模型的目标变量进行

doi:10.3772/j.issn.2095-915x.2018.03.014

划分，如表1所示。

如图2所示，为有效提升评价结果的精确度，基于电力业务行为的用电客户价值评价模型根据上述目标变量的划分，综合考虑预付能力、用电需求、欠费风险、安全风险、诈骗风险、增容趋势、减容趋势、销户风险八个子预测模型的评价结果，最终形成用电客户的总体评价。其中，每个预测结果的价值评分可从预测结果为电力企业带来的营收贡献和运营成本来衡量，而营收贡献和运营成本则可由电力企业的财务报表获得，进而得到每个子预测模型对用电客户的预测结果评分，最终对其进行求和得到用电客户的价值评分。

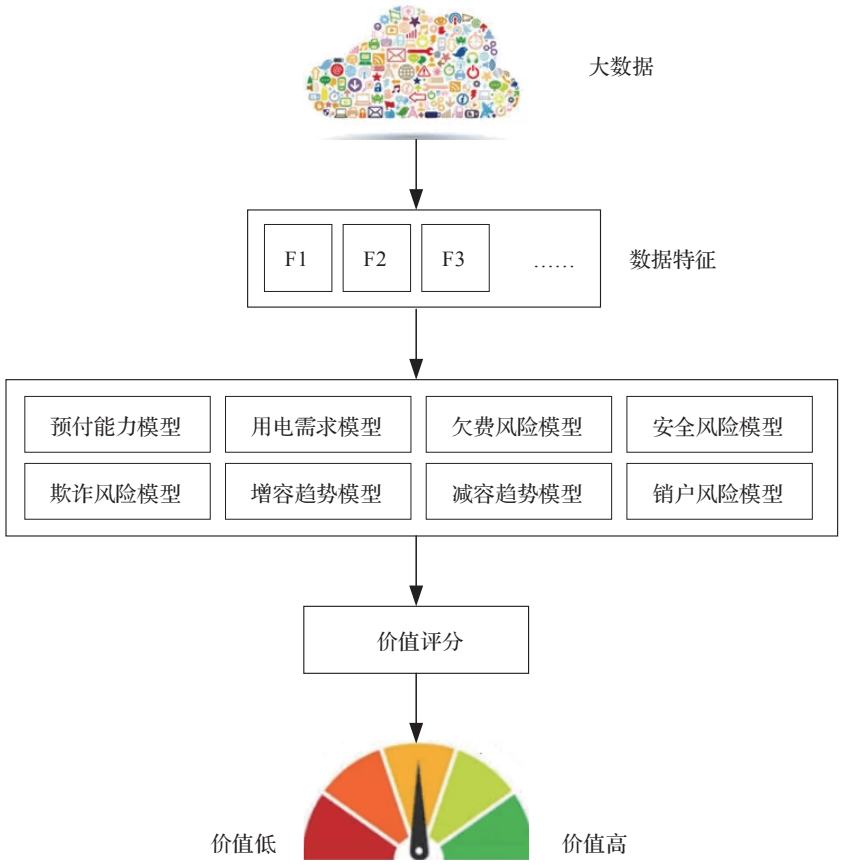

图 2　基于电力业务行为的用电客户价值评价模型

3 结合数据筛选的半监督随机森林算法

预测模型的准确率是上述价值评价模型的关键。电力行业数据存在明显的类别、数据不平衡问题，某些类别的用户群数量非常少。传统的分类算法在此类问题下，泛化能力较差，而随机森林算法作为一种集成方法，在处理高维度数据和对抗噪声干扰方面表现良好，且具

有很强的泛化能力。本文采用无标记样本生成及获取与半监督的随机森林算法相结合[14]，来构建电力业务行为预测模型。

3.1 无标记样本的生成和获取

对抗生成网络(Generative Adversarial Network，GAN)是当前机器学习领域的研究热点，通过判别器和生成器的构建及对抗训练，生成与真实样本具有相同分布的新样本，而后，将此生成样本加入到训练集中，以有效解决训练集的类不平衡问题。传统的对抗生成网络中普遍存在由于生成器梯度消失而导致的训练失效、生成器梯度不稳定和样本多样性不足的问题[15-16]，针对上述问题，本文采用对抗生成网络的衍生模型Wasserstein GAN(以下简称WGAN)来进行无标记样本的生成[17-18]。

使用WGAN生成无标记样本不可避免会遇到以下难点：1)如何设置WGAN的迭代训练停止条件，使得生成的样本数据对半监督预测模型精度产生积极影响；2)如何降低生成数据中真实标签以外类别的数据对半监督模型的影响[19-20]。为解决此难点，本文训练出一种指示分类器，即使用WGAN生成的无标记样本和真实样本作为训练集训练出的分类器，对WGAN生成的无标记样本集进行筛选后，再进行半监督学习。在此过程中，当指示分类器的分类准确率达到最大时，对应的WGAN的迭代次数为最佳迭代次数n，迭代停止，而后使用指示分类器对生成的样本进行筛选。

指示分类器的构建方法基于Label Smoothing Regularization(LSR)理论[21]，该理论最早在二十世纪80年代被提出，通过在输出y中添加噪声，实现对模型的约束，进而降低模型过拟合，它最早被用于有监督学习中。通常有监督学习模型使用0或1的方式对输出的真实分布进行编码，0表示一个样本不属于该类别，1则表示属于该类别，形式化表达为：

$$q(m)=\begin{cases}0 & m\neq y_i\\ 1 & m=y_i\end{cases},\ m\in Y=\{y_1,y_2,y_3,\dots,y_n\}$$

（公式1）

其中，$q(m)$为真实概率分布，Y为有监督学习中的标签。LSR技术主要是为了缓解标签不够平滑而导致的过拟合问题，该方法让模型判断一个样本属于某个类别的概率不是100%，不属于某个类别的概率也不是100%，即

$$q(m)=\begin{cases}\dfrac{\varepsilon}{K} & m\neq y_i\\ 1-\varepsilon+\dfrac{\varepsilon}{K} & m=y_i\end{cases},\quad m\in Y=\{y_1,y_2,y_3,\dots,y_n\}$$

（公式2）

其中，ε是一个超参数，K表示标签集合Y的个数。这样使得模型获得一个新的标签分布。LSR技术通过降低模型对于它的预测的置信度来避免模型对样本的预测严重偏离真实的情况，从而避免过拟合现象的产生。本文在LSR的基础上，为生成的无标记样本数据设置一个虚拟的标签分布，为

$$q(m)=\frac{1}{K},\quad K=n+1,m\in Y$$

（公式3）

其中，n表示真实样本数据中标签集合$Y=\{y_1,y_2,\dots,y_n\}$的个数，Y表示标签集合$Y=\{y_1,y_2,\dots,y_n,y_{generated}\}$，$y_{generated}$表示该类样本是生成的，$m$表示标签集合$M$中的标签，$K$表示标签集合$Y$的个数，$q(m)$表示某样本属于

doi:10.3772/j.issn.2095-915x.2018.03.014

标签m的概率。同时，真实样本标签转化为$y=\{y_1,y_2,\ldots, y_n,y_{generated}\}=\{y_1,y_2,\ldots,y_n,0\}$，代表真实样本数据属于$y_{generated}$标签的概率为0，其余标签分布不变。

本文对生成的无标记样本设置统一的标签分布是基于假设：1)一个训练良好的WGAN必然具有足够好的样本多样性而不是集中与某几个“安全样本”附近；2)生成的样本大多由真实样本数据中的特征重新组合生成得到，假设生成的样本不能完全属于某个真实样本标签$Y=\{y_1,y_2,\ldots,y_n\}$；3)由LSR理论假设对于所有类别采用统一分布可以有效降低过拟合。由于多层神经网络模型具有较强的拟合能力，因此，借助神经网络算法可以得到一个由真实样本和生成样本训练得到的多分类器，使用交叉熵损失值作为多层神经网络模型的损失函数。WGAN模型用于生成半监督学习的无标记样本数据的算法如表2、表3所示，其中，算法2是算法1的子流程。

表 2　训练对抗生成网络的衍生模型 WGAN

算法1 训练WGAN

输入：WGAN参数：真实样本集X；真实样本集数据维度X_dim；噪声数据维度z_dim；隐藏层节点数h_dim；迭代次数集合$epochs=\{e_1,e_2,\ldots,e_n\}$；学习率$\alpha$；梯度惩罚系数$\lambda$；$\varepsilon\in[0,1]$为随机数；生成样本集大小$m_{fake}$；

输出：WGAN模型、生成样本X_g

步骤一：初始化判别器D和生成器G的系数w_D和w_G，以及噪声生成器z;

步骤二：迭代训练生成器G和判别器D，直到满足停止条件。

while 迭代次数$it\neq e_i$ do

for t=1,…,n do

for i=1,…,(m/$batch\_size$) do

抽样$\{x_j\}_{j=0}^{batch_size}\sim P_r$，$\{z_j\}_{j=0}^{batch\_size}\sim p(z)$，生成一个随机数$\varepsilon\in U[0,1]$

$x_g=G(z)$

$x_{gp}=\varepsilon x+(1-\varepsilon)x_g$

$L_i=D(x_g)-D(x)+\lambda\left(\left\|\nabla x_{gp}D(x_{gp})\right\|_2-1\right)^2$

$w_D\leftarrow Adam\left(\nabla_{w_D}\frac{1}{batch\_size}\sum_{i=1}^{batch\_size}L_i,w_D,\alpha,\beta_1,\beta_2\right)$

end for

end for

抽样得$\{z_j\}_{j=0}^{batch_size}\sim p(z)$

$w_G\leftarrow Adam\left(\nabla w_G\frac{1}{batch\_size}\sum_{i=1}^{batch\_size}-D(G(z)),w_G,\alpha,\beta_1,\beta_2\right)$

end while

步骤三：输出生成大小为m_i的生成样本X_g

表 3　无标记样本生成方法

算法2 无标记样本获取算法

输入：指示分类器网络参数：真实样本集X及标签集$y=\{y_1,y_2,\cdots,y_n,y_g\}$，$n$为真实标签类别数，$y_g$表示生成样本类别标签；测试样本集$X_{test}$和标签$y_{test}$；迭代次数training_epoch；隐藏层节点数h_dim；学习率β；批量大小batch_size；

真实样本标签大小K;

输出：

无标记样本集S_U^*、最优指示分类器、WGAN最优迭代次数

步骤一：使用真实样本集X预训练得到基准分类器C_{S_L}，并计算分类准确率$acc_{C_{S_L}}$；

步骤二：计算不同迭代次数下的$acc_{C_{S_l+S_u}}$

for e_i in epochs：

WGAN迭代训练e_i次，生成大小为m_f的生成样本X_g；

设置生成样本集的标签y的概率都为$1/K+1$；

设置真实样本集X的标签为：

$$y=\{y_1,y_2,\cdots,y_n,y_{generated}\}=\{y_1,y_2,\cdots,y_n,0\}$$

初始化指示分类器的神经网络参数w；

while 迭代次数不满足training_epoch：

使用训练集$X\cup X_g$最小化指示分类器损失函数：

$$loss=-(1-Z)\log(p(y))-\frac{Z}{K}\sum_{k=1}^{K}\log(p(m))$$

，当样本为真实样本时，Z=0，否则Z=1;

$$w\leftarrow Adam\left(\nabla w\frac{1}{batch\_size}\sum_{i=1}^{batch\_size}loss,w,\beta,\beta_1,\beta_2\right)$$

end while

计算测试样本集X_{test}指示分类器的$acc_{C_{S_l+S_u}}$，并将（e_i，$acc_{C_{S_l+S_u}}$）存入集合L；

end for

步骤三：获取集合中L中$acc_{C_{S_l+S_u}}$最大时的e_i，重新迭代训练WGANe_i次；

步骤四：利用指示分类器$C_{S_l+S_u}$筛选输出标签$y_{generated}$概率低于$1/K+1$为无标记样本集S_U^*。

3.2　结合数据筛选的半监督随机森林算法

对于结构化数据，随机森林算法在抵抗噪声和分类性能上都具有明显优势，且受类不平衡问题的影响较小。本文在传统的随机森林算法的基础上，引入半监督学习中的协同训练思想，在Co-Forest[22]模型的决策树集合的基础上，通过不断筛选出无标记样本中的噪声，来保证半监督模型性能的提升。本部分主要阐述数据筛选方法，即如何将分歧较大的“高质量”数据筛选出来用于训练，以及半监督的随机森林模型。

根据Bagging算法思想[23]，在大小为m、数据维度为d的带标签训练数据集L上采用Bootstrap的方法[24]有放回的选取m个样本，并随机抽取不同的d个列，共进行n轮迭代得到n

doi:10.3772/j.issn.2095-915x.2018.03.014

个不同的训练集，训练后得到n个决策树分类器$T=\{t_1,t_2,\ldots,t_n\}$。使用T_i表示除t_i外其余n-1个决策树分类器的集合，称为组合分类器。在4.1部分，通过WGAN生成大量无标记的数据集，并通过指示分类器筛选出能提高分类器性能的数据，最终得到无标记样本数据集U，大小为M。本文定义每一个无标记样本x的分类置信度$w_{i,t}$，为组合分类器T_i中对该无标记样本x分类结果的一致性。采用组合分类器T_i的分类一致性超过一定的阈值θ，给无标记样本打上伪标签y，并加入到新标记数据集U_i'中。在得到无标记样本及其伪标签之后，在训练的过程中不断筛选和处理某些置信度较低的或错误标记的样本，以提高分类器的泛化能力。噪声指的是训练样本集中不属于真实样本期望分布的数据。

半监督的分类模型的关键是对新标记样本的判断，当新标记样本的分类错误率较高时，将会严重影响半监督分类模型的分类能力。数据筛选的思想是基于半监督思想的平滑假设。因此，本文提出的数据筛选方法主要针对以下类型的新标记样本数据x进行筛选：1)样本数据x的K个近邻里标签与x相同的数量少于k；2)与样本数据x距离较近的样本标签与x不一致；3)无标记样本中的离群点。前两种类型违反了半监督学习的基本假设，第三种类型的数据对于分类器的性能提升效果不大，反而如果无标记样本中的离群点与标记样本分布的距离太大容易引入大量的噪声，从而降低分类器的泛化能力，因此，本文将以上三种类型的无标记样本成为可疑标记样本。

第一种可疑标记样本通过K近邻(KNN)算法进行筛选，其中，参数K和k按照Generalized Editing设定的条件$(K+1)/2<k<K$来设置。本文采用参数$k=7$，$k=5$。对于第二种类型的可疑标记样本x，其标签为y，定义K个近邻的样本集合为$Q=\{q_1, q_2, q_3,\ldots, q_K\}$，在$Q$中标签和伪标签与样本$x$相同的数据点集合为$Q_S$，标签不同的样本点集合为$Q_D$。根据$Q$中的样本点与$x$的不一致性和距离，定义筛选条件为：

$$\sum_{z\in Q_S}\left(\varpi_{x,z}*\left(\sum_{y'\in Y}\left(w_{x,y'}-w_{z,y'}\right)^2\right)\right)<\sum_{u\in Q_D}\left(\varpi_{x,u}*\left(\sum_{y'\in Y}\left(w_{x,y'}-w_{z,y'}\right)^2\right)\right)$$

$$\varpi_{x,z}=e^{\frac{d_{x,z}^2}{\sigma^2}}, \quad \text{（公式4）}$$

其中，$d_{x,z}$表示样本x与样本z的距离，σ表示样本x与Q中所有点距离的平均值。$w_{x,y'}$表示样本点x的标签为y'时分类器集合T的预测一致性。当样本x满足公式的时候被筛选出来。对于第三种可疑标记样本，该类样本与真实标记的样本整体距离较大，采用样本到真实标记样本的距离之和对其进行筛选，定义偏离度为：

$$DP(x)=\sum_{z\in L}\varpi_{x,z} \quad \text{（公式5）}$$

其中，$DP_{\max}$为新标记样本集中最大的偏离度，每轮迭代的筛选条件为：

$$DP_{\max}(x)<(1-epoch_{select}\cdot\beta)\cdot DP_{\max} \quad \text{（公式6）}$$

其中，$epoch_{select}$为当前数据筛选迭代次数，β为距离衰减率，默认为0.001。数据筛选出来的可疑样本，不作为训练样本。半监督的随机森林算法具体步骤如表4所示。

4 实验与分析

4.1 实验数据准备及预处理

本文以冀北地区2014年1月至2016年6月期

间大工业用户的月用电量数据、月无功电量数据、月线损电量数据、用电检查记录、违约用电记录、容量变更数据、缴费记录、95588拨打记录作为实验数据，构建预测模型。通过预测冀北地区大工业用户2015年1月到2016年1月期间的容量变更风险来验证半监督随机森林的分类性能。预测模型的目标变量为2015年1月、2015年7月、2016年1月用电客户是否有减容风险，其中，有减容风险的样本大小为12405条，无减容风险的样本大小为20963条。为了提高模型收敛速度、提高精度，前期已对训练样本进行数据归一化预处理。

本文基于Python 语言在Tensorflow框架上快速搭建上述基于电力业务行为的用电客户价值评价预测模型。此处半监督随机森林预测模型分为两部分：一是WGAN生成大量的无标记样本，并使用指示分类器对无标记样本进行筛选，以提高半监督分类器的分类性能；二是基于协同训练思想的半监督随机森林分类器，在训练过程中，对无标记样本进行筛选并处理以提高分类性能。

表 4　半监督随机森林

算法3 半监督随机森林

输入：标记训练样本L，大小为m_l，无标记训练样本U，大小为m，预测一致度阈值θ，随机森林中决策树分类器数量n，列抽样数量d，距离衰减率β，筛选条件K和k，对于筛选出来的数据的处理方式data_select；

输出：半监督的随机森林分类模型T

步骤一：初始化所有决策树分类器和组合分类器

通过有放回的从L中抽取n个大小为m_l，数据维度为d的训练集；

训练得到n个决策树分类器$T=\{t_1,t_2,\ldots,t_n\}$；

初始化组合分类器$e_{i,0}=0.5$，新标记样本权重和$W_0=\sum_{j=0}^{m_0}1$；

步骤二：对于每一个组合分类器，迭代以下过程，直到迭代次数到达或者决策树不再变化。

for t_i in T:

计算组合分类器T_i在有标记样本L上的错误率$e_{i,l}$，令新标记集合为$U_{i,l}=\phi$。

end for

if $e_{i,l}>e_{i,l-1}$:

使用Bootstrap方法在无标记样本U抽取样子集，大小为m，通过T_i计算每一个无标记样本x_u的置信度，且当置信度大于置信阈值θ时将样本复制到新标记集合$U_{i,l}$中，即$U_{i,l}=U_{i,l}\cup x_u$。

end if

for t_i in T:

if $\varepsilon_{i,l}W_{i,l}<\varepsilon_{i,l-1}W_{i,l-1}$:

令it = 0，$\varepsilon_0W_0=\varepsilon_{i,l}W_{i,l}=\varepsilon_{it}W_{it}$;

do :

it = it + 1

对新标记集合进行数据筛选，并进行处理

重新计算$\varepsilon_{it}W_{it}$

while $\varepsilon_{it}>\varepsilon_{it-1}$;

使用筛选后的数据重新训练分类模型；

else：

直接使用未进行数据筛选的训练集$L\cup U_{i,l}$重新训练分类模型

end if

end for

步骤三：输出模型$T(x)$

分类过程：n个决策树分类器的预测结果通过少数服从多数的投票机制产生。

doi:10.3772/j.issn.2095-915x.2018.03.014

4.2 实验结果与分析

本文在MNIST手写数字数据集[25]上对上WGAN生成用于半监督学习的无标记样本的方法进行验证。生成器G和判别器D都采用两层隐藏层的神经网络模型。不同迭代次数下，WGAN生成的无标记样本训练的指示分类器的准确率不同，如表5所示。

表 5　不同迭代次数下 WGAN 对分类器性能的提高

迭代次数	baseline	20000	30000	40000	50000	55000	60000	70000
准确率	64.7%	65.7%	70.2%	75.0%	80.7%	83.6%	84.3%	84.0%

由表5可知，随着迭代次数的不断增加，指示分类器的分类准确率不断提升，最后达到一个峰值，然后保持稳定。因此，采用WGAN迭代60000次后生成的样本作为无标记样本，利用指示分类器进行筛选后，结果如图3所示。

4.2.1 生成样本对指示分类器的影响

采用有标记样本16000条，其中，有减容风险的1000条，无减容风险的15000条，分别采用由WGAN训练迭代10000次后生成的不同样本集大小的无标记样本，加伪标签后，加入到训练集中，来验证由WGAN生成的无标记样本集大小对指示分类器的影响，其分类结果的准确度如表6所示。

由表6可知，当往指示分类器的训练集中添加带伪标签的无标记样本时，指示分类器的分类性能有了一定程度的提高，最大提升百分比为2.64%。当无标记样本集规模较小时，指示分类器的泛化能力较弱，随着无标记样本规模的增大，指示分类器的泛化性能中图提升，但当无标记样本的数量增加到一定量后，其对指示分类器分类性能的提高作用逐渐减小。出现此现象的原因是由于WGAN生成的无标记样本虽然与用于训练的真实样本分布类似，但因真实样本的特征互相组合，导致样本多样性大大增加，产生了大量不属于原有标签的样本数据。因此，伴随着无标记样本数量的增多，过拟合现象也越来越严重，致使指示分类器性能提升减小。

(a)　(b)　(c)

图 3　(a) 为真实样本、(b) 为生成样本、(c) 筛选后的样本

表 6　WGAN 生成的无标记样本集大小对指示分类器的影响

样本大小	0	10000	12000	13000	14000	15000
准确率	73.96%	74.18%	75.62%	76.60%	76.83%	76.07%

4.2.2 无标记样本对半监督随机森林模型的影响

将经过指示分类器筛选后的不同样本集大小的无标记样本加入到训练集中，用于训练半监督随机森林分类模型，无标记样本大小分别设置为5000，10000，15000，20000，25000。起初设置的半监督随机森林模型迭代终止条件是决策树分类器集合中各个分类器都不再发生变化，但经实验发现，达到该停止条件需要花费较长的时间，因此，采用设置迭代次数来进行取代，这里，迭代次数设置为20000次。此外，我们对每个训练集进行10次随机采样，并用其分别对模型进行训练，将每次得到的分类模型准确率、有减容风险用电客户查全率的平均值作为结果，以避免样本分布不均匀给模型的带来不良影响。

表 7　无标记样本集大小对半监督随机森林模型的影响

训练集	准确率	查全率
baseline	76.63%	59.39%
标记样本+0无标记样本	75.38%	52.43%
标记样本+5000无标记样本	77.01%	71.11%
标记样本+10000无标记样本	78.42%	70.56%
标记样本+15000无标记样本	77.92%	70.99%
标记样本+20000无标记样本	78.76%	72.21%

由表7可以看出，在类别不平衡的情况下，传统的随机森林模型比没有使用无标记样本的半监督随机森林分类性能要高。但随着无标记样本的增多，半监督的随机森林模型比随机森林模型在准确率上有一定程度的提高。对于有减容风险的用电客户的预测，在使用无标记样本进行训练前的准确率较低，但在加入生成样本后，模型的查全率有了明显的提高。这意味着，Wasserstein GAN模型一定程度上找到了减容用电客户的数据分布模式，生成的样本集跟真实的样本集的差异性相对较小。实验结果表明，在无标记样本不断增多的情况下，对模型的分类效果的提高并不明显，但由于采用了数据筛选方法，无标记样本中的错误标记样本对模型泛化能力的影响大大降低。

4.2.3 有标记样本对半监督随机森林模型的影响

本部分主要通过实验验证将更多有标记样本集加入到训练集中对模型带来的影响。设置基准的训练集共16000条，逐渐增加的有标记样本的数量分别为1000、2000、3000、4000、5000、6000条，分别构成新的不同大小的训练集，类别比例1:1，对传统的随机森林模型进行训练。对于本文提出的半监督的随机森林模型，则去掉新增加的有标记样本的标记，作为无标记样本加入训练集中进行训练，模型使用的参数保持不变，模型分类准确率如表8所示。由表8可知，在传统的随机森林模型中，真实的有标记样本的加入对于模型的分类能力有明显的提高，且提高的幅度比半监督随机森林模型高2.57%。但是半监督随机森林模型在加入的

doi:10.3772/j.issn.2095-915x.2018.03.014

真实的有标记样本较少时，分类性能的提高不如使用WGAN生成并经过筛选的无标记样本加入训练集所能带来的分类性能的提高。由此可知，WGAN生成并经过筛选的无标记样本一定程度拟合了真实的样本分布，并且在类不平衡的情况下，为数量较少的类提供了更为多样性的样本，在半监督学习中更好地辅助寻找模型的决策边界。

表 8 真实样本对模型的影响

训练集	随机森林	半监督随机森林
baseline	76.63%	75.38%
原训练集+1000标记样本	77.75%	75.47%
原训练集+2000标记样本	79.30%	76.62%
原训练集+3000标记样本	80.99%	78.53%
原训练集+4000标记样本	81.52%	79.24%
原训练集+5000标记样本	82.14%	79.67%
原训练集+6000标记样本	82.01%	79.20%

5 评价模型在冀北电力公司的实际应用

本文提出的用电客户价值评价模型需要对用电客户的预付能力、用电需求、欠费风险、违约风险、减容风险、安全风险、增容趋势、销户风险进行预测。由于上述目标变量的发现周期较长，且发生的次数较少，因此，本文统一对用电客户未来六个月的业务行为进行预测，并设置预测点为2015年1月、2015年7月、2016年1月。由于本文中使用的欠费数据和安全隐患数据只有2016年1月到2016年9月这段范围内的，所以只将这两个模型的预测点设置为2016年1月。由此得到的减容风险、增容趋势、违约风险、欠费风险四个训练集的类别分布如表9所示。其中，标记为“1”的目标变量与价值负相关、标记为“0”的目标变量与价值正相关，同时模型的预测结果为0或1。同时，所有模型的训练集与测试集以9：1的比例进行划分。最终得到的各个预测模型的预测能力如表10所示。由于具体的模型权重系数为企业商业机密，因此本文采用平权对用电客户进行评价。取价值评分最高的前10%作为高价值用户，冀北五个地级市在2016年7月的高价值用户的分布如图4所示。

表 9 训练集的类别分布

	标记为“1”	标记为“0”
减容趋势	12405	20963
增容风险	1360	32008
违约风险	2414	30954
欠费风险	7080	6940

表 10 四类用电行为预测模型的分类能力

模型	准确率
减容风险	82.01%
增容趋势	76.30%
违约风险	74.29%
欠费风险	80.43%

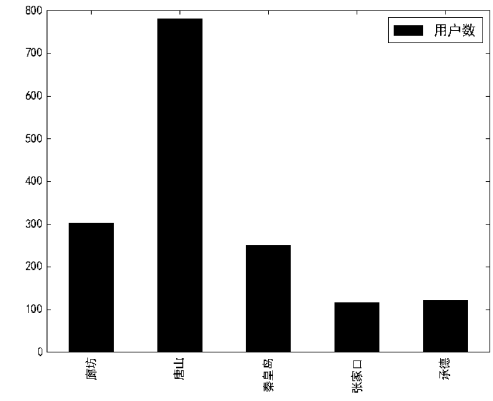

图 4 高价值用户在各地级市的分布

由于唐山市汇集了冀北地区最多的重工业用户，特别是特大型的钢铁厂，是高耗电用电客户，因此，相对唐山市高价值客户会比其他地市多。经过与各地市下属单位的业务专家进行鉴定，所筛选出来的高价值用户均为电力企业中的“一级客户”和“重点客户”，此外，所选出的客户均具有用电需求大、欠费风险低、三方用户等特征，符合电力企业对于高价值用户的需求。

6 结语

用电客户价值评价是电力企业的工作重点之一，本文所提出的基于电力业务行为的用电客户价值评价模型，其核心就是将价值评价问题转化为分类问题，而后采用了对抗生成网络、半监督随机森林等机器学习算法，构建电力业务行为预测模型，有效解决电力营销数据中类不平衡的问题，并对用电客户的行为实现较为精准的预测。本文为用电客户的价值评价提供了一条新的思路，并对当前针对用电客户价值评价遇到的问题提供了一种新的解决方案，具有较强的商业价值和应用潜力。

参考文献

[1] Asadinejad A, Rahimpour A, Tomsovic K, et al. Evaluation of residential customer elasticity for incentive based demand response programs[J]. Electric Power Systems Research, 2018, 158:26-36.

[2] Lawi A, Wungo S L, Manjang S. Identifying irregularity electricity usage of customer behaviors using logistic regression and linear discriminant analysis[C]. International Conference on Science in Information Technology. 2017:552-557.

[3] 林森, 欧阳柳. 基于大数据理论的电力客户标签体系构建[J]. 电气技术, 2016, 17(12):98-101.

[4] Kim T, Lee D, Choi J, et al. Extracting Baseline Electricity Usage Using Gradient Tree Boosting[C]. IEEE International Conference on Smart City/socialcom/sustaincom. IEEE, 2015:734-741.

[5] Rathod R R, Garg R D. Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data[J]. International Journal of Energy Sector Management, 2017, 11(2):295-310.

[6] 周辛南, 谢枫, 傅军,等. 基于德尔菲法的电力大客户综合价值评价模型[J]. 电测与仪表, 2016, 53(15A):174-177.

[7] 赵晓东. 面向大数据的供电企业电力营销服务体系构建分析[J]. 内蒙古电力技术, 2016, 34(4):29-31.

[8] 陈缨, 卢思瑶, 沈焱,等. 电力客户价值评估与管理的实证研究[J]. 西南民族大学学报(人文社科版), 2017, 38(9):130-133.

[9] 杨慧霞, 邓迎君, 刘志斌,等. 含有历史不良数据的电力负荷预测研究[J]. 电力系统保护与控制, 2017, 45(15):62-68.

[10] 唐静, 李瑞轩, 黄宇航,等. 基于多维特征分析的月用电量精准预测研究[J]. 电力系统保护与控制, 2017, 45(16):145-150.

[11] 李婉华, 陈宏, 郭昆,等. 基于随机森林算法的用电负荷预测研究[J]. 计算机工程与应用, 2016, 52(23):236-243.

[12] Dai X B, Huang Y S. Research of Electricity Enterprise Credit Assessment Combining Fuzzy Multiple Attribute Decision Making with Improved Analytic Hierarchy Process -Taking Baoding Electricity Supply Company as Example[C]. 21ST International Conference on Industrial Engineering and Engineering Management, Zhuhai, China, Nov 1-2, 2014.

[13] Du W, Zhou C, Zheng B, et al. Comprehensive

doi:10.3772/j.issn.2095-915x.2018.03.014

Credit Evaluation Model of Electricity Customer Based on the Changing Trend of Credit[C]. International Conference on Information Science and Cloud Computing Companion. IEEE, 2014:536-542.

[14] Zhou Z H, Li M. Tri-Training: Exploiting Unlabeled Data Using Three Classifiers[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(11):1529-1541.

[15] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.

[16] Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J]. Computer Science, 2015.

[17] Arjovsky M, Chintala S, Bottou L. Wasserstein Generative adversarial networks [EB/OL]. [2018-03-21]. http://proceedings.mlr.press/v70/arjovsky17a.html .

[18] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks[C]. The 5th International Conference on Learning Representations, Toulon, France, 2017.04.24-26.

[19] Cui S, Jiang Y. Effective Lipschitz constraint enforcement for Wasserstein GAN training[C]. IEEE International Conference on Computational Intelligence and Applications. IEEE, 2017:74-78.

[20] Masaki S, Eiichi M, Shunta S. Temporal Generative Adversarial Nets with Singular Value Clipping[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, Oct 22-29, 2017.

[21] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]. Computer Vision and Pattern Recognition. IEEE, 2016:2818-2826.

[22] Li M, Zhou Z H. Improve Computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2007, 37(6):1088-1098.

[23] Dong X S, Qian L J, Huang L. A CNN based bagging learning approach to short-term load forecasting in smart grid[C]. 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, San Francisco, USA, Aug 4-8, 2017.

[24] Ezekwesili R, Shahzad M K, Baboli, A, et al. Optimizing Bootstrap method to improve forecasting accuracy in business jet spare parts supply chains[C]. 4th International Conference on Optimization and Applications, Mohammedia, Morocco, April 26-27, 2018.

[25] The MNIST Database of Handwritten Digits[EB/OL]. [2018-01-23]. http://yann.lecun.com/exdb/mnist/.