参赛队员姓名：杨逸驰　徐光梓

中学：杭州外国语学校

省份：浙　　江

国家/地区：中华人民共和国

指导教师姓名：徐亦飞　邓水光

论文题目：基于监督学习的窃电行为识别

基于监督学习的窃电行为识别

杨逸驰，徐光梓

摘要

近年来，我国电网等基础设施建设快速发展，电力行业的管理者面临日益突出的窃电问题，用户窃电行为给供电企业经济效益稳定增长和社会发展等方面造成负面影响。传统反窃电手段要求执法人员逐户排查电表是否完整、接线是否正确，需要耗费大量的人力物力财力。随着电网信息化建设的完善，用户的用电数据记录较为完整。因此，如何挖掘用户数据行为、识别异常用户，已经成为电网大数据研究中的热门话题。

本文的研究目的在于利用监督学习识别异常用电用户，为供电企业稽查工作提供参考，有效的降低用电运营成本 。本文主要贡献包括以下三个方面：

1）提出了多视角的用电特征，降维、归一化的用户数据，为描述用户用电行为提供参考。本文首先使用单属性指标（均值、差值、变异系数、极差、标准差） 和多属性指标（余弦相似性和皮尔逊系数）对用户用电量（月、季度、年、节假日、工作日等）进行特征表述，接着利用主成分分析（PCA）等技术进行维数约简，最后使用归一化和过抽样技术形成特征数据集。

2）研究了多种数据分类方法，探究最优的参数配置，为识别用户用电行为提供模型。本文分别对支持向量机、BP 神经网络、随机森林、XGBoost 分类方法及其相关的参数进行研究，得到性能最佳的分类器。

3）比较了多种数据分类方法，使用不同的指标评估模型，为判别用户用电行为提供指导。本文在真实的用电数据集上，应用不同的分类模型，在多个指标上进行性能评估，并依据不同的应用场景推荐最优模型。

关键字：窃电，监督学习，用电行为

Abstract

With the rapid development of infrastructure including power grids, managers in power industry these days are faced with an increasingly severe problem of theft of electricity. Theft of electricity has negative effects on many socioeconomic aspects, including impacting stable growth of economic for power enterprises and social development. The traditional anti-theft means require officers checking the integrity of kilowatt-hour meter and the correctness of wiring house by house, which requires enormous manpower and material resources. With the advancement of information collecting technology, power enterprises now possess relatively complete database of power consumption. As a result, performing data mining on existing database and identifying abnormal users has become a hot topic in the field of information technology.

The purpose of this paper is to identify abnormal users with machine learning algorithms, providing power enterprises with means to detect theft of electricity at lower cost. The contributions of this paper are listed as follows:

1) Propose various features, providing means to describe users' behaviors. First, single variable features (mean, difference, coefficient of variation, range, standard deviation) and double variable features (cosine similarity, Pearson product-moment correlation coefficient) are calculated based on user (month, season, year, holiday, workday) power consumption. Then principal component analysis is used to perform dimensionality reduction on the dataset. The dataset is finally scaled and oversampled to form the feature dataset.
2) Study various classification algorithms, optimize hyperparameters, providing models to detect theft of electricity. In this paper, the mechanism behind support vector machine, BP neural network, random forest and XGBoost is studied and the hyperparameters are optimized.
3) Compare and evaluate different classifiers, provide suggestions for real-world application. In this paper, different classifiers are evaluated with different metrics and best classifier is recommended based on real-world dataset and different application scenarios.

Keywords: theft of electricity, surpervised learning, user behavior

目录

1. 引言

随着中国电力行业的快速发展，窃电的问题日益突出。窃电者使用非法手段少交、不交电费，严重扰乱市场秩序，影响供电企业的正常盈利。除此之外，窃电行为引起的供电不足影响正常用户，妨碍居民正常生活。窃电过程中乱接电线、改装电表易引起安全隐患，可能导致火灾等意外[1]。

尽管窃电技术不断升级，窃电手法愈发隐蔽，用户的非法窃电行为不可避免地从统计数据中暴露出来。随着应用于电力系统的科学技术不断发展，今天供电企业有较为完整的用电记录数据库，而如何从用电数据中识别窃电行为成为反窃电信息化的热点和难点[2, 3]。

如图 1 所示，本文提出了基于监督学习的窃电行为模型，其主要包括输入、系统和输出三个部分。在系统部分，本文提出了有效地用户用电行为特征和高效的用户行为分类模型，为供电企业稽查工作提供参考，有效的降低了运营成本 。

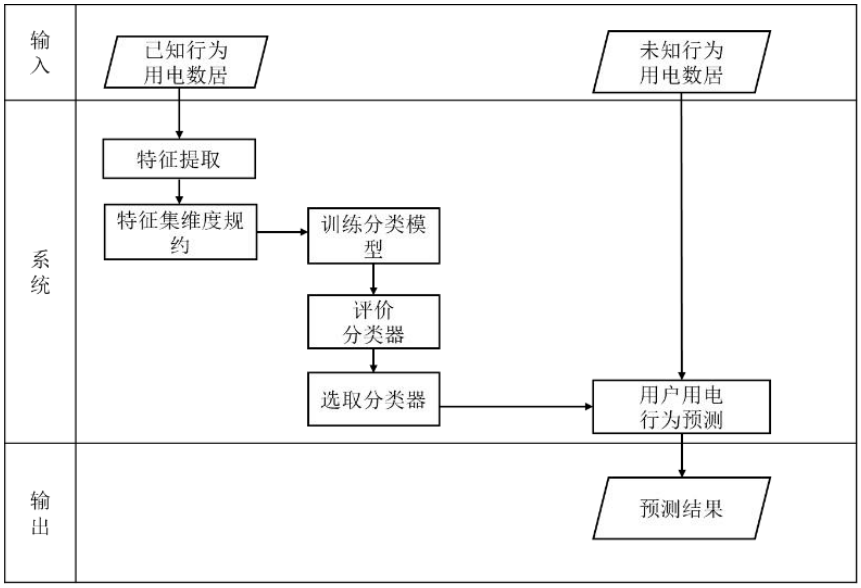

图 1. 基于监督学习的窃电行为识别模型示意图
Fig. 1. Theft of electricity detection model based on supervised learning

本文主要贡献包括以下三个方面：

1）提出了多视角的用电特征，降维、归一化的用户数据，为描述用户用电行为提供参考。本文首先使用单属性指标（均值、差值、变异系数、极差、

标准差） 和多属性指标（余弦相似性和皮尔逊系数）对用户用电量（月、季度、年、节假日、工作日等）进行特征表述，接着利用主成分分析（PCA）等技术进行维数约简，最后使用归一化和过抽样技术形成特征数据集。

2）研究了多种数据分类方法，探究最优的参数配置，为识别用户用电行为提供模型。本文分别对支持向量机、BP 神经网络、随机森林、XGBoost 分类方法及其相关的参数进行研究，得到性能最佳的分类器。

3）比较了多种数据分类方法，使用不同的指标评估模型，为判别用户用电行为提供指导。本文在真实的用电数据集上，应用不同的分类模型，在多个指标上进行性能评估，并依据不同的应用场景推荐最优模型。

本文的内容安排如下，第 2 章提出不同用户用电特征，对用户行为进行描述，通过主成分分析进行数据维约，并列举模型的评估指标；第 3 章分析了分析用电行为的不同分类算法；第 4 章讨论了实验分析及结果，包含了数据集预处理，模型参数优化和实验结果分析；第 5 章对本文进行总结，并讨论了下一步的工作；第 6 章是参考文献。

2. 用电行为描述

2.1 用电行为特征提取

用户用电行为可以使用代表性的用电特征进行描述。首先从原始数据集中提取 N 个用电用户 H 天的用电数据，则每个用户的用电数据可表示为 H 维向量 $x_n=\{x_n^{(h)},h\in Dates\}$，其中 Dates 是数据集中所有日期的集合，所有用户组成的数据集为$X=\{x_n,n=1,2,3,\ldots,N\}$。对数据集X，本文使用下列统计指标对用电特征进行表述，其中a_i为数据项，T、K 为数据向量。用户用电特征主要包括单属性特征、双属性特征和符合属性特征。单属性特征指的是单一数据项的统计值，双属性特征指的是两个数据向量的统计值，复合属性特征指的是单一或两个属性的混合统计值。

$$A_{(变异系数)}=\frac{\sqrt{\sum_{i=1}^{n}(a_i)^2}}{\frac{1}{n}\sum_{i=1}^{n}a_i} \quad (1)$$

$$B_{(标准差)}=\sqrt{\sum_{i=1}^{n}(a_i)^2} \quad (2)$$

$$C_{(极差)}=Max(a_i)-Min(a_i) \quad (3)$$

$$D_{(均值)}=\frac{1}{n}\sum_{i=1}^{n}a_i \quad (4)$$

$$E_{(余弦相似性)}=\frac{\sum_{i=1}^{n}T_i\times K_i}{\sqrt{\sum_{i=1}^{n}(T_i)^2}\times\sqrt{\sum_{i=1}^{n}(K_i)^2}} \quad (5)$$

$$F_{(皮尔逊相关系数)}=\frac{cov(T,K)}{\sigma T\sigma K} \quad (6)$$

2.1.1 单属性特征

1) 三年月用电量的特征

三年月用电量的变异系数$A_1^{(1)}(A_{1\text{-}2014}^{(1)},A_{1\text{-}2015}^{(1)},A_{1\text{-}2016}^{(1)})$，标准差$B_1^{(1)}(B_{1\text{-}2014}^{(1)},B_{1\text{-}2015}^{(1)},B_{1\text{-}2016}^{(1)})$，极差$C_1^{(1)}(C_{1\text{-}2014}^{(1)},C_{1\text{-}2015}^{(1)},C_{1\text{-}2016}^{(1)})$，均值$D_1^{(1)}(D_{1\text{-}2014}^{(1)},D_{1\text{-}2015}^{(1)},D_{1\text{-}2016}^{(1)})$，共计特征 3+3+3+3= 12 个。

2) 节假日和非工作日的特征

节假日电量和每天用电量比值的变异系数$A_2^{(1)}$，标准差$B_2^{(1)}$，极差$C_2^{(1)}$，均值$D_2^{(1)}$；工作日电量与非工作日电量比值的变异系数$A_2^{(2)}$，标准差$B_2^{(2)}$，极差$C_2^{(2)}$，均值$D_2^{(2)}$，共计特征 4+4=8 个。

3) 每年前 5 个月与后 5 个月的差值的指标

每年前 5 个月与后 5 个月的差值$H_8^{(1)}(H_{8\text{-}2014}^{(1)},H_{8\text{-}2015}^{(1)},H_{8\text{-}2016}^{(1)})$，3 年前 5 个月与后 5 个月的差值的变异系数$A_8^{(1)}$，标准差$B_8^{(1)}$，极差$C_8^{(1)}$，均值$D_8^{(1)}$，共计特征 3+4=7 个。

2.1.2 双属性特征

1) 三年月电量关联的特征

两年月电量余弦相似性$E_3^{(1)}(E_{3\text{-}2014\&2015}^{(1)},E_{3\text{-}2015\&2016}^{(1)})$，皮尔逊相关系数$F_3^{(1)}(F_{3\text{-}2014\&2015}^{(1)},F_{3\text{-}2015\&2016}^{(1)})$，皮尔逊相关系数差值$G_3^{(1)}=\left|E_{3\text{-}2014\&2015}^{(1)}-E_{3\text{-}2015\&2016}^{(1)}\right|$，共计特征 2+2+1=5 个。

2) 三年节假日电量关联的特征

两年节假日电量余弦相似性$E_4^{(1)}(E_{4\text{-}2014\&2015}^{(1)},E_{3\text{-}2015\&2016}^{(1)})$，皮尔逊相关系数$F_4^{(1)}(F_{4\text{-}2014\&2015}^{(1)},F_{4\text{-}2015\&2016}^{(1)})$，皮尔逊相关系数差值$G_4^{(1)}=\left|E_{4\text{-}2014\&2015}^{(1)}-E_{4\text{-}2015\&2016}^{(1)}\right|$，共计特征 2+2+1=5 个。

3) 三年季度电量关联的特征

两年季度电量余弦相似性$E_5^{(1)}(E_{5\text{-}2014\&2015}^{(1)},E_{5\text{-}2015\&2016}^{(1)})$，皮尔逊相关系数$F_5^{(1)}(F_{5\text{-}2014\&2015}^{(1)},F_{5\text{-}2015\&2016}^{(1)})$，皮尔逊相关系数差值$G_5^{(1)}=\left|E_{4\text{-}2014\&2015}^{(1)}-E_{4\text{-}2015\&2016}^{(1)}\right|$，共计特征 2+2+1=5 个。

4) 用户和行业用电情况关联的特征

行业用电情况指所有正常用户的平均用电的各项指标，可以使用该指标对窃电用户行为进行比对识别。

a) 用户和行业月用电量余弦相似性$E_6^{(1)}(E_{6\text{-}2014}^{(1)},E_{6\text{-}2015}^{(1)},E_{6\text{-}2016}^{(1)})$，皮尔逊相关系数$F_6^{(1)}(F_{6\text{-}2014}^{(1)},F_{6\text{-}2015}^{(1)},F_{6\text{-}2016}^{(1)})$

b) 用户和行业季度用电量余弦相似性$E_6^{(2)}(E_{6\text{-}2014}^{(2)},E_{6\text{-}2015}^{(2)},E_{6\text{-}2016}^{(2)})$，皮尔逊相关系数$F_6^{(2)}(F_{6\text{-}2014}^{(2)},F_{6\text{-}2015}^{(2)},F_{6\text{-}2016}^{(2)})$

c) 用户和行业节假日用电量余弦相似性$E_6^{(3)}(E_{6\text{-}2014}^{(3)},E_{6\text{-}2015}^{(3)},E_{6\text{-}2016}^{(3)})$，皮尔逊相关系数$F_6^{(3)}(F_{6\text{-}2014}^{(3)},F_{6\text{-}2015}^{(3)},F_{6\text{-}2016}^{(3)})$

共计特征 6+6+6=18 个。

2.1.3 复合属性特征

用户的用电数据斜率指的是用户相邻日期的用电量斜率，可以有效的反映用户用电量的波动情况。斜率差异指的是斜率和行业斜率的差异，可以有效的反映用户与正常用户的差异情况。

a) 用电斜率的变异系数$A_7^{(1)}$，标准差$B_7^{(1)}$，极差$C_7^{(1)}$，均值$D_7^{(1)}$

b) 斜率差异的变异系数$A_7^{(2)}$，标准差$B_7^{(2)}$，极差$C_7^{(2)}$，均值$D_7^{(2)}$

c) 用电斜率和行业斜率余弦相似性$E_7^{(3)}$,皮尔逊相关系数$F_7^{(3)}$

共计特征 4+4+2=10 个

2.2 特征集维度归约

主成分分析（PCA）是一种数据降维算法，其原理为利用线性拟合将分布在多个维度的高维数据投射到有限维度上，保留贡献大的主成分，忽略贡献小的主成分，可以有效的保留原始变量反映的信息目的[4, 5]。

设样本有 n 个变量$x_1, x_2, x_3 \ldots\ldots x_n$，令向量$X=[x_1\ x_2\ x_3 \ldots\ldots x_n]$，用$PC_1$, PC_2,$PC_3 \ldots\ldots PC_m$表示原变量的 m 个主成分，令主成分 i 的载荷(loading) $A_i^T=[x_{i1}\ x_{i2}\ x_{i3} \ldots\ldots x_{in}]$，则有

$$\begin{cases} PC_1=A_1^TX \\ PC_2=A_2^TX \\ PC_3=A_3^TX \\ \quad\ldots\ldots \\ PC_m=A_m^TX \end{cases} \tag{7}$$

2.3 模型分析指标

模型的性能需要使用指标进行评估。本文采用准确率（accuracy）、精确率（precision）和召回率（recall）评价分类器的分类效果。在分类结果中，TP 为将

正类预测为正类的数目，FN 为将正类预测为负类的数目，FP 为将负类预测为正类的数目，TN 为将负类预测为负类的数目数，则

$$\text{precision}=\frac{TP}{TP+FP} \tag{8}$$

$$\text{recall}=\frac{TP}{TP+FN} \tag{9}$$

$$\text{accuracy}=\frac{TP+TN}{TP+FN+FP+TN} \tag{10}$$

在本文中，准确率为用电用户正确分类的比例，精确率是预测的窃电用户为真实窃电用户的比例，召回率是真实窃电用户被判定为窃电用户的比例。为了精准地识别窃电用户，本文主要对精确率进行评估。

3. 用电行为识别

用户行为识别主要指的是使用多种不同的分类器对用户行为进行识别，本文讨论了支持向量机、BP 神经网络、随机森林及 XGBoost 等模型及其模型参数。

3.1 支持向量机

支持向量机(Support Vector Machine)可以有效的解决小样本、非线性及高维模式的模式识别，并能够推广应用到函数拟合等其他机器学习问题中[6]。如图 2 所示，在线性可分的情况下，支持向量机通过寻找到一个超平面 H，使离 H 最近的正负样本刚好分别落在 H1 和 H2 上，并且使函数间隔（H1 和 H2 的距离）最大，以达到将样本准确分为正负两类的目的。

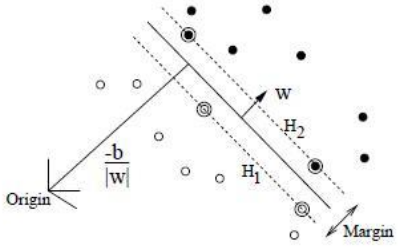

图 2. SVM 构建过程

Fig. 2. Construction of SVM

转换为优化问题，SVM 问题可以转换为如下的条件极值问题：

$$\begin{cases} \dfrac{\min\|w\|^2}{2} \\ \text{s.t.} \quad y_i(w\cdot x_i+b)-1\geq 0 \end{cases} \tag{11}$$

其中 w 为权重系数。通常，建立 SVM 分类器时会使用核函数对数据进行某非线性变换，将其映射到更高维的空间，分离在低维空间中不易区分的数据样本 。常见的核函数包括’rbf’和’linear’。

3.2 BP 神经网络

BP(back propagation)神经网络是是一种按照误差逆向传播算法训练的多层反馈神经网络，是目前应用最广泛的神经网络[7]。该算法通过隐藏输入与输出的具体关系，使用数据集训练达到接受输入后产生接近期望值的输出的功能。BP 神经网络模型的实现包含信号的前向传播和误差的反向传播两个主要过程。其具体的工作流程如下图 3 所示：

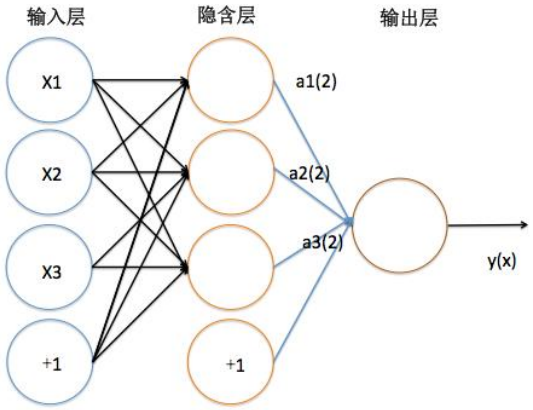

图 3. BP 神经网络的构建

Fig. 3. Construction of BP neural network

假设建立一个包含 h 个隐含层的 BP 神经网络，其各隐含层节点数按前向顺序分别记为 m1,m2,…,mh，各隐含层输出分别记为 y1,y2,…,yh，各层权值矩阵分别记为 W1,W2,…,Wh,Wh+1，则各层权值调整公式如下。

输出层：

$$\Delta w_{jk}^{h+1}=\eta\delta_{h+1}^{k}y_{j}^{h}=\eta(d_{k}-o_{k})o_{k}(1-o_{k})y_{j}^{k} \quad (j=0,1,2,\ldots,mh;k=1,2,\ldots,l) \tag{12}$$

第 h 隐含层：

$$\Delta w_{ij}^{h}=\eta\delta_{j}^{h}y_{i}^{h}\text{-}1=\eta\left(\sum_{k=1}^{l}\delta_{k}^{0}w_{jk}^{h+1}y_{j}^{k}(1\text{-}y_{j}^{k})y_{i}^{h}\text{-}1\right) \ (i=0,1,2,\ldots,m_{(h-1)};j=1,2,\ldots,l) \tag{13}$$

信号的前向传播过程中，输入向量从输入层进入，经隐含层逐层处理，在输出层产生输出。若神经网络的误差达到可以接受的范围（或者达到了预先设定的学习次数），则结束算法；若实际输出与期望输出偏差较大，则进入误差反向传播的过程， 基于偏差修正每层各单元的权值。前向传播和反向传播交替进行，直至偏差下降至预定范围。构建 BP 神经网络模型的过程中，学习速率和隐含层的单位数是影响模型性能的重要参数。

3.3 随机森林

随机森林是一种集成学习算法，首先对数据集进行行采样，从 N 个输入样本中有放回地抽取 N 个样本，然后对数据进行列采样，从 M 个特征中，选择 m（m<M）个特征，最终用采样得到的数据建立决策树并重复该过程多次，得到 k 颗不同决策树组成随机森林。新数据的分类结果由所有决策树投票决定。随机森林对多元共线性数据不敏感，结果对缺失数据和非平衡的数据比较稳健[8]。

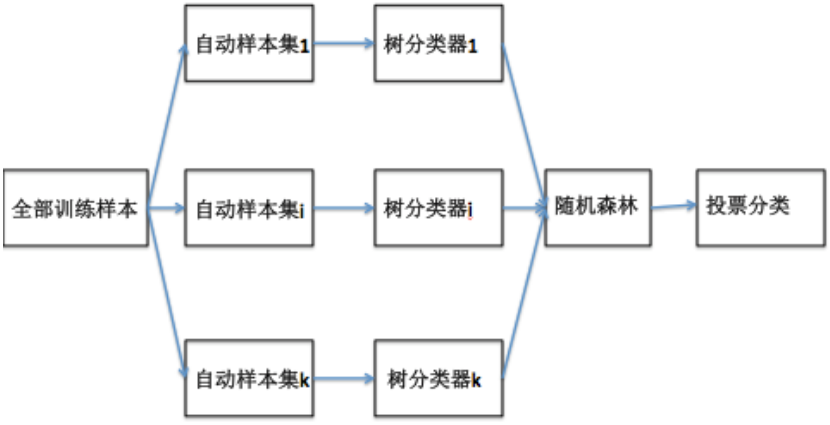

图 4. 随机森林的构建过程

Fig. 4. Construction of Random Forest

如图 4 所示，随机森林是由多个独立决策树{h(X, θk),k=1,2,...,N} 组成的组合分类器。输入特征集 X 后，每个决策树分别作出判断，并以每树一票投票得到随机森林分类器的最终分类结果。最大树深度和决策树数量是随机森林模型构建过程中两个很重要的参数。

3.4 XGBoost

XGBoost 是开源的梯度提升决策树（GBDT）实现，克服了传统梯度提升决策树难以并行化，模型复杂度高的缺点，在提升决策树的基础上引入带有正则项的目标函数，同时支持不同分类器和自定义损失函数，具有并行处理、模型复杂度低，过拟合可能性低等优点[9-11]。

不同于 GBDT，XGBoost 定义了（式 14）函数，用作监督学习的目标函数。

$$Obj=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i\right)+\sum_{i=1}^{n}\Omega(f_k) \tag{14}$$

其中正则项控制模型复杂度，防止发生过拟合。

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_i^2 \tag{15}$$

XGBoost 通过 additive training 降低目标函数值，即训练时每一步保持之前的分类与回归树（CART）集不变，加入一棵新的 CART 使目标函数值减少量最大。通过二阶泰勒展开进行近似并忽略常数后，得到第 t 步的目标函数如下（式 16）：

$$Obj^{(t)}=\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \tag{16}$$

其中 $g_i=\partial_{\hat{y}_i^{(t-1)}} l(y_i,\hat{y}_i)$；$h_i=\partial^2_{\hat{y}_i} l(y_i,\hat{y}_i)$，可见 XGBoost 可以使用任何一个二阶可导的函数作为损失函数。由上代入正则项可得到式 17：

$$Obj^{(t)}=\sum_{j=1}^{T}\left[G_j w_j+\frac{1}{2}\left(H_j+\lambda\right)w_i^2\right]+\gamma T \tag{17}$$

对于给定的t，该目标函数是一个关于w_j的二次函数，由此可以求得如下最小值（式 18）：

$$obj^{*}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j}{H_j+\lambda}+\gamma T \tag{18}$$

得到CART叶子节点分割产生左、右两个叶子节点时目标函数减少量，即分割的增益：

$$Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma \tag{19}$$

生成一棵CART时遍历可能的分割方法，用上式计算出增益，选取增益最大的分割。最后，生成的所有CART通过投票的方式确定给定特征对应的类别。同时，XGBoost也采用了类似随机森林的列抽样，可以提高训练速度并避免过拟合。

在 XGBoost 中，学习率、决策树数量、决策树深度等参数影响了模型的性能。

4. 实验结果与分析

4.1 用户数据集

本数据集是某地用户 2014 年 1 月到 2016 年 10 月以日为单位的用电数据，窃电和正常用户数据比例为 3799 比 40419。数据集包含两张数据表，表 1 包含用户用电数据，表 2 包含用户用电行为标识（是否窃电）。

表 1. 原始数据集 1
Table 1. Original dataset 1

字段	含义	示例
CONS_NO	用户 ID	0E13DA5C01DEEFA7E11B5DD0BAA1542A
DATA_DATE	日期	2014/1/1
KWH_READING	当日结束时电表示数	626.64
KWH_READING1	当日开始时电表示数	625.75
KWH	当天用电量	0.89

表 2. 原始数据集 2

Table 2. Original dataset 2

字段	含义	示例
CONS_NO	用户 ID	0E13DA5C01DEEFA7E11B5DD0BAA1542A
LABEL	标识（0：正常；1：窃电）	0

数据集中的记录依据时间排序，共有 1034 天数据。每个用户关联多条用电记录，即存在单一用户 ID 具有日期不同的多条记录，用来表示同一用户不同日期的用电情况。除去数据集中的无效值和空值（无用电数据），共得到有效用户 42026 人。该数据集描述了两类用户用电行为：0 类（正常行为），该用户行为正常，为合法用户；1 类（窃电行为），该用户行为窃电，为非法用户。0 类用户共有记录 38264，1 类用户共有记录 3762，样本比例为 10.2:1。

4.2 数据集预处理和交叉验证方法选取

4.2.1 数据集预处理

数据集的预处理包括数据修补、数据划分和数据归一化。

1）数据修补

用户数据包括用户 ID、日期、当日结束时电表示数、当日开始时电表示数、当天用电量等数据项。由于每日用户用电量是当日结束时电表示数与当日开始时电表示数的差值，所以可以删除冗余项当日结束时电表示数与当日开始时电表示数。

删减后的用户数据包括用户 ID、日期及用户日用电量。对用户数据按用户 ID 和日期排序，排序完成后得到每个用户的按日期排列的用电量列表。

2）数据划分

数据划分包括特征提取和样本抽样。

特征提取：根据第 2 章提出的用户行为描述方法，生成包括不同用户用电特征的 2 个数据集。第 1 个数据集的大小为42026×70。样本数目为 42026 个，样本特征为 70 个，包括三年月用电量的指标、节假日和非工作日的指标、每年前 5 个月与后 5 个月的差值的指标、两年月电量关联的指标、两年节假日电量关联的指标、两年季度电量关联的指标、用户和行业用电情况关联的指标、用电斜率和斜率差异的指标；第 2 个数据集使用统计学中的一些度量值生成特征，分别是日用电量平均值、日用电量最小值、日用电量最大值、日用电量方差和日用电量中位数。

样本抽样：由于数据集中两类样本分布不均衡，若直接用来训练分类器易使最终训练得到的分类器分类效果差，因此对 1 类进行过抽样（Oversampling），使最终数据集中两类记录数相等。

3）数据归一化

因为数据特征的量度不同，为了分类更加精确，需要对数据进行归一化处理。归一化公式见式 20，Fea为特征列，Fea_{norm}为归一化后的特征列。

$$Fea_{norm}=\frac{Fea\text{-}Min(Fea)}{Max(Fea)\text{-}Min(Fea)} \tag{20}$$

4.2.2 交叉验证方法的选取

交叉验证是一种统计学上将数据样本切割成较小子集的实用方法，其具体操作是使用一个数据子集训练模型，用剩余的子集检验模型泛化到独立数据集的能力。

简单交叉验证中数据集被按比例（例如 80%，20%）随机分为训练集和测试集，分别用于训练模型和测试模型。该方法的缺点是仅利用训练集的数据训练模型，无法充分利用数据。为克服数据利用不充分的问题，本文采用分层 K 折（Stratified K-Folds）交叉验证，将除测试集以外的数据分为 K 折，用分层的方法使每一折两类（即正常和窃电）数量大致相等。训练分类器时每次选取 1 折作为验证集，剩余 K-1 折作为训练集并训练和验证分类器。重复该过程 K 次，获得平均准确率、精确率和召回率，用来对模型的整体性能进行评估[12]。

4.3 模型参数分析

4.3.1 支持向量机参数分析

支持向量机有两个重要的参数：惩罚因子 C 和 gamma，这两个参数与核函数相关，影响数据处理的准确性和召回率。

1) 惩罚因子 C

C 一般可以选择为：10t , t=[-4,4]， 数值与例惩罚程度相关，惩罚程度越大，模型过拟合可能性越高；惩罚程度低会导致模型准确性和召回率过低。在其他参数都相同的情况下，本文使用网格搜索参数确定方法，当 C=1000 时，取得最佳的模型性能（accuracy= 0.592,precison= 0.757 ; recall= 0.270 ; time= 31091.43s）。

2) gamma

gamma 是‘rbf’核函数的重要参数，决定了数据映射到新的特征空间后的分布。本文使用网格搜索方法分别对‘rbf’和‘linear’核函数进行分析，由实验结果得知，在其他参数都相同的情况下，选择‘linear’作为 kernel 的准确度和召回率最高。

4.3.2 BP 神经网络参数分析

BP 神经网络模型中有两个重要参数：学习速率（learning_rate_init）和隐含层的单位数（hidden_layer_sizes），对于收敛速度和训练结果影响很大。

1) 学习速率（learning_rate_init）

BP 网络神经算法是基于误差-修正学习的，学习速率是 BP 算法的重要部分，其大小对收敛速度和训练结果影响很大。学习速率太小，学习速度太慢，模型收敛消耗时间过长；学习速率太大，可能导致模型振荡或发散。为得到最优的学习速率，本文设计了下述的参数评估实验。

在隐含层单位数和隐含层数相同的情况下，选择了以下几个学习速率来寻找最佳的学习速率：0.0005,0.001,0.003,0.005,0.007,0.01，实验结果如表 3 和图 5 所示。

表 3. 不同 learning_rate 和 hidden_layer_sizes 下 BP 模型精确度
Table 3. BP Model precision table for different learning_rate and hidden_layer_sizes

hidden_layer_sizes/ 隐含层单位数	learning_ rate/ 学习速率	Accuracy/ 准确率	Precision/ 精确率	Recall/ 召回率	Time(s)/ 时间
85	0.0005	80.26%	79.55%	81.75%	215.52
85	0.001	82.58%	81.81%	83.85%	172.06
85	0.003	82.90%	81.83%	84.60%	160.63
85	0.005	79.92%	78.74%	82.24%	170.63
85	0.007	76.82%	73.32%	84.41%	142.14
85	0.01	74.08%	71.80%	79.79%	128.15
30	0.003	75.10%	73.68%	78.22%	88.70
50	0.003	79.26%	79.05%	79.91%	118.30
70	0.003	81.95%	81.10%	83.51%	135.99
100	0.003	83.58%	81.94%	86.13%	231.29
110	0.003	84.90%	83.11%	87.71%	222.24

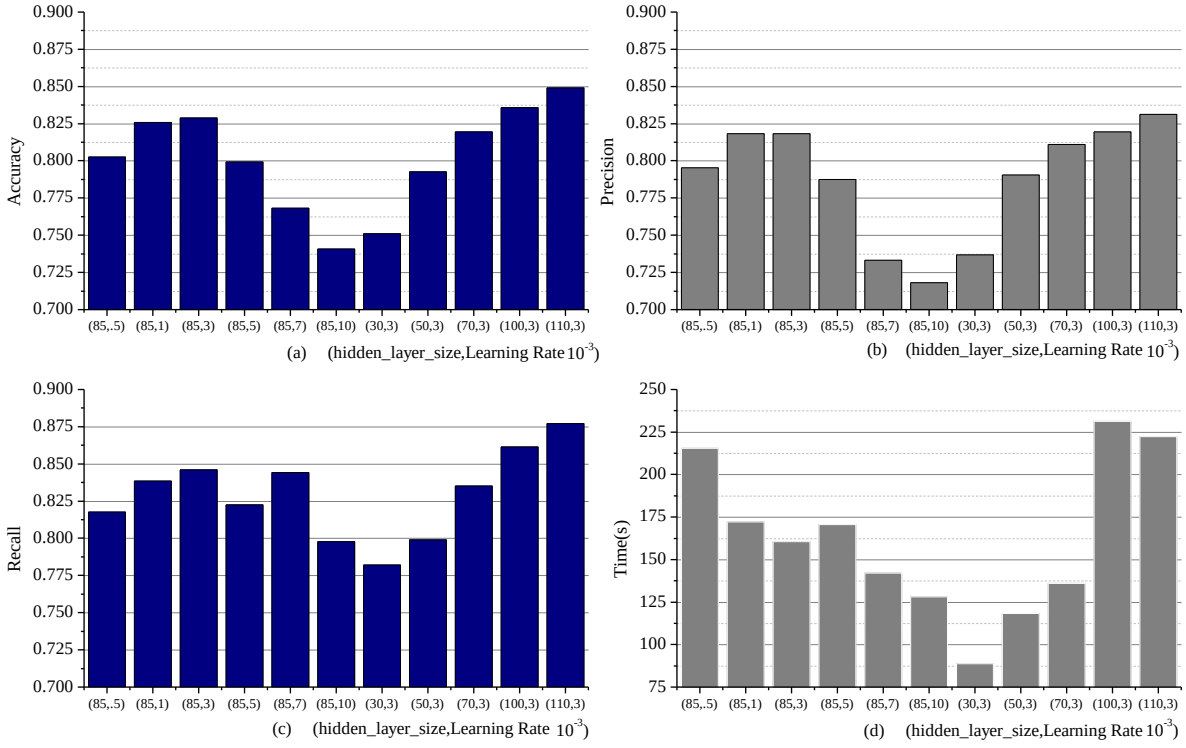

图 5. BP 神经网络不同隐含层单位数和学习速率下实验结果
Fig. 5. BP Model precision graph for different learning_rate and hidden_layer_sizes

如表 3 可以得出，当学习速率为 0.003 或 0.001 时，数据处理得到的准确率和召回率较高，其它学习速率导致准确率大幅度下降，但当学习速率为 0.001 时所需的时间比学习速率为0.003所需时间长，且准确率没有明显的提高，所以选择0.003作为最优的学习速率。

2) 隐含层的单位数（hidden_layer_sizes）

隐含层的单位数会影响收敛速度和训练结果，实验预测的准确性会随着隐藏层单位数的增加而提高，但同时也会层数增多运行时间明显增长。当隐含层的单位数到达某个值之后，实验预测的准确性增幅很小。为此，为了获得最优的隐含层的单位数，本文设计了如下实验。

在其它参数都为默认值的情况下(隐含层层数选择为 2 层)，选择了以下隐含层单位值来寻找最佳的隐含层单位数：30,50,70,85,100,110。如图5所示，当隐含层单位数小于 85 时，实验预测的准确率和召回率会随着隐含层单位数的增加有明显提高；而当隐含层单位数在 85 的基础上继续增加时，虽然准确率和召回率有所提高，但运行时间明显增长，因此，本次实验选择 85 作为最优的隐藏层单位数。

4.3.3 随机森林参数分析

最大深度（max_depth）和决策树数量（n_estimators）决定了随机森林模型的性能，本文设计了相关的参数评估实验搜寻最佳的模型参数。

1) 最大深度（max_depth）

决策树的最大深度会影响模型运行速度及模型准确性：如果决策树的最大深度太小，模型准确性较差，但决策树的最大深度达到某个阈值，模型准确性的提升不明显，运行时间变长。为寻找最佳的最大深度，如表 4 和图 6 所示，在固定其它参数的情况下，本文使用了一系列的最大深度对模型性能进行测试。

从表 4 和图 6 可以看出，当其他参数固定的情况下，决策树的最大深度为默认值 None 时，模型的准确性和召回率最高，因此，本文选择 None 作为决策树最优的最大深度。

表 4. 不同 n_estimators 和 max_depth 下 RF 模型的精确度

Table 4. RF Model precision table for different n_estimators and max_depth

n_estimators/ 决策树数量	max_depth/ 最大深度	Accuracy/ 准确率	Precision/ 精确率	Recall/召回率	time/时间(s)
10	10	82.03%	80.35%	84.80%	17.75
20	10	83.11%	81.63%	85.47%	32.64
30	10	83.22%	81.49%	86.00%	46.14
40	10	83.13%	81.49%	86.00%	64.45
50	10	83.83%	82.29%	86.05%	77.65
40	20	97.75%	96.54%	86.21%	86.40
40	30	98.66%	98.28%	99.05%	90.84
40	None	98.68%	98.32%	99.05%	91.85
30	30	98.57%	98.13%	99.06%	68.31
20	30	98.55%	98.06%	99.04%	45.95
10	30	98.40%	97.79%	99.05%	22.37
5	30	82.03%	80.35%	84.80%	17.75
10	None	83.11%	81.63%	85.47%	32.64

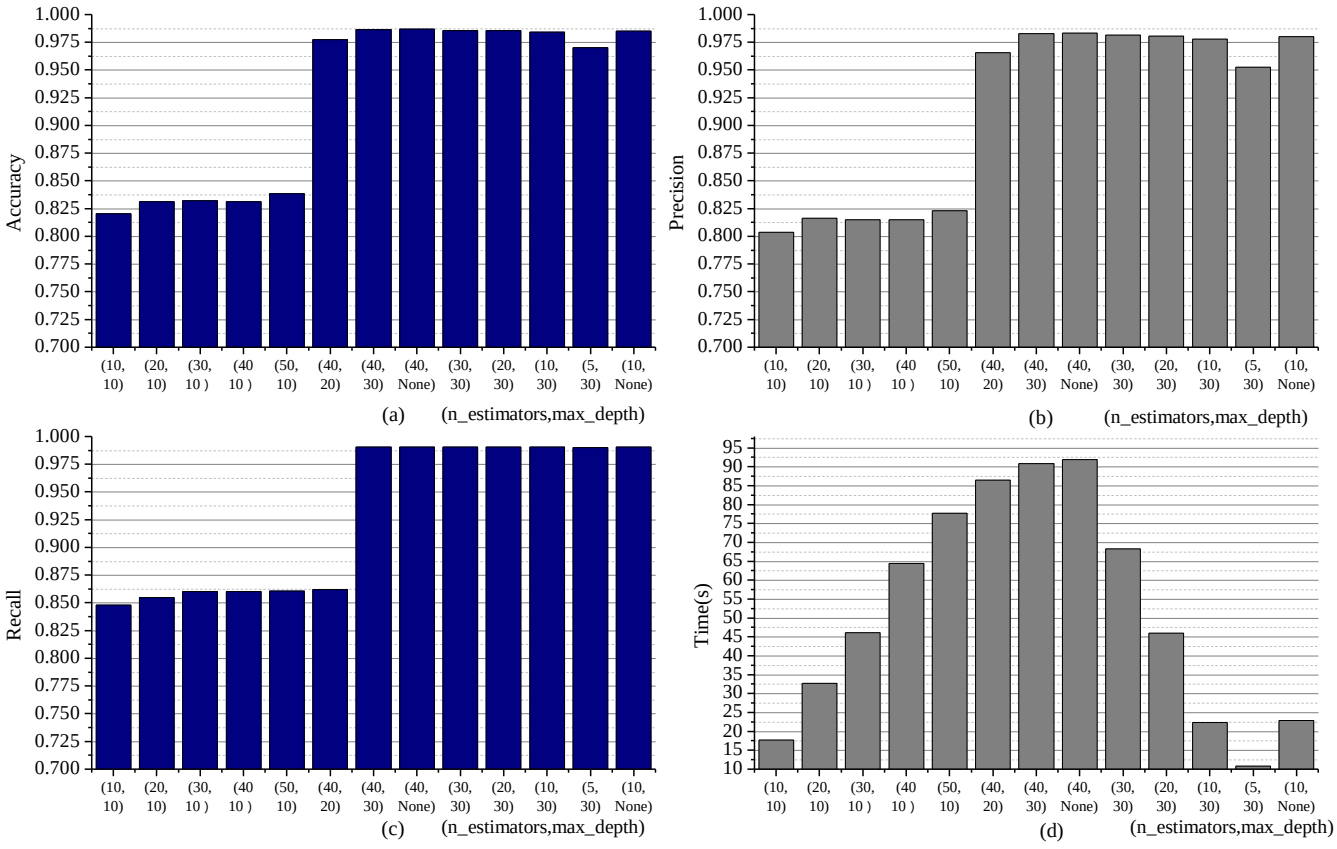

图 6. RF 随机森林不同决策树数量和最大深度下实验结果

Fig. 6. RF Model precision graph for different n_estimators and max_depth

2) 决策树数量（n_estimators）

决策树的数量也会影响模型的准确性和运行速度，当决策树的数量过小，会模型的准确度和召回率过低；当决策树的数量过大，虽然能够保证模型的准确度，但是会增加模型的运行时间。为了寻找最优参数，在其他参数都为默认值的情况下（由上一个实验可得决策树的最佳深度为 30），从多个隐含层单位数（10,20,30,40,50）寻找最佳的决策树数量。

如图 6 所示，当其他参数都相同的情况下，决策树的数量从 10 开始，准确度和召回率基本不变，而决策树的数量为 10 时，所用的时间最少。所以本次实验选择 10 作为最优决策树的数量。

4.3.4 XGBoost 参数分析

超参数是在开始学习过程之前设置的参数，而不是通过训练得到的参数数据。超参数的设置直接影响模型的性能，因此对超参数进行调节至关重要。在 XGBoost 模型中，主要有以下重要超参数：控制学习速率和迭代次数的 earning_rate 和 n_estimators；控制 CART 生长的 max_depth 、min_child_weight 和 gamma；控制行列采样的 subsample 和 colsample_bytree；正则项系数 reg_lambda。

1) learning_rate 和 n_estimators；

learning_rate 是 XGBoost 算法生成 CART 时节点权重的缩减率，n_estimators 为集成学习模型中 CART 的数量，与模型性能的优劣密切相关。通常情况下，learning_rate 越低，模型的鲁棒性越好，但运算时间会增长。为了平衡运行时间和模型性能之间的矛盾，因此，在本次实验中，首先将 learning_rate 设置为较高值，快速的完成其它超参数的调节。接着， 固定其它超参数，不断降低 learning_rate 及提高 n_estimators 数量，寻求可控时间内使模型性能最优的最优参数。

2) max_depth 、min_child_weight 和 gamma

XGBoost 分类器由许多 CART 构成，max_depth 、min_child_weight 和 gamma 是控制 CART 生成的重要超参数。max_depth 控制 CART 最大深度，

min_child_weight 控制 CART 分裂时叶节点最小权重，gamma 为允许分裂的增益最小值。增大 min_child_weight 和 gamma 会使模型更加保守；增大 max_depth 会提高模型复杂度，这意味着更容易过拟合。为了避免过拟合，提高模型泛化能力，本次实验中通过网格搜索的方法，找到 max_depth 、min_child_weight 和 gamma 的最优值 9,1,和 0（实验结果见表 5、6 和图 7）。

表 5. 不同 max_depth 、min_child_weight 下模型分类精确度表

Table 5. Model precision table for different max_depth and min_child_weight

max_depth	min_child_weight	accuracy	precision	recall
3	1	87.62%	84.25%	92.52%
3	3	86.85%	83.65%	91.59%
3	5	86.29%	83.23%	90.88%
5	1	95.42%	92.49%	98.87%
5	3	94.87%	91.64%	98.75%
5	5	94.29%	90.87%	98.48%
7	1	96.81%	94.73%	99.14%
7	3	96.29%	93.81%	99.12%
7	5	95.87%	93.11%	99.08%
9	1	97.14%	95.33%	99.14%
9	3	96.57%	94.29%	99.14%
9	5	96.25%	93.74%	99.13%

表 6. 不同 gamma 下模型分类精确度表

Table 6. Model precision table for different gamma

gamma	accuracy	precision	recall
0	97.14%	95.33%	99.14%
0.1	96.96%	95.00%	99.14%
0.2	96.68%	94.51%	99.11%
0.3	96.37%	94.00%	99.06%
0.4	96.33%	93.94%	99.06%

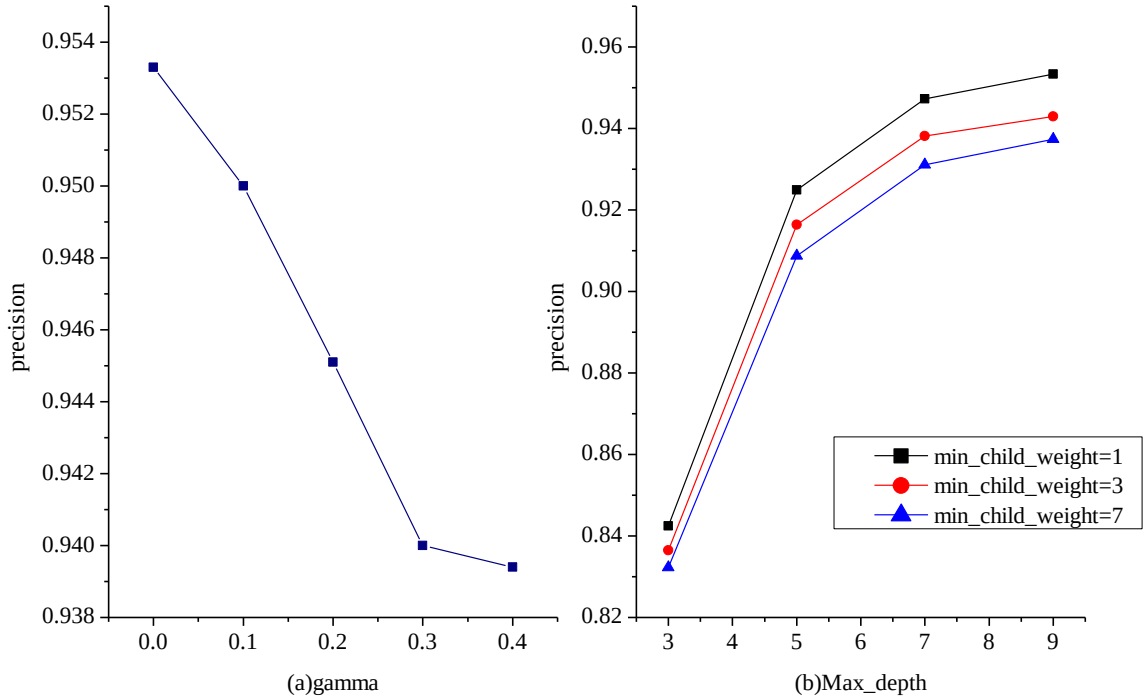

图 7. 不同 max_depth 、min_child_weight 和 gamma 下模型分类精确度曲线

Fig. 7. Model precision curve for different max_depth ,min_child_weight and gamma

3) subsample 和 colsample_bytree

XGBoost 允许对数据集利用行抽样和列抽样以避免过拟合和降低运算量。subsample 是对数据集进行行抽样的比例，colsample_bytree 是对特征进行列抽样的比例。为了提高分类器的训练效果同时降低运算量，实验中通过网格搜索的方法选取 subsample 和 colsample_bytree 的最优值 0.9 和 0.8（实验结果见表 7 和图 8）。

表 7. 不同 subsample 和 colsample_bytree 下模型分类精确度表

Table 7. Model precision table for different subsample and colsample_bytree

subsample	colsample_bytree	accuracy	precision	recall
0.6	0.6	96.16%	93.58%	99.13%
0.6	0.7	96.16%	93.58%	99.13%
0.6	0.8	96.34%	93.90%	99.12%
0.6	0.9	96.34%	93.90%	99.12%
0.7	0.6	96.32%	93.88%	99.10%
0.7	0.7	96.32%	93.88%	99.10%
0.7	0.8	96.40%	94.00%	99.12%
0.7	0.9	96.40%	94.00%	99.12%
0.8	0.6	96.28%	93.82%	99.09%

续表 7

subsample	colsample_bytree	accuracy	precision	recall
0.8	0.7	96.28%	93.82%	99.09%
0.8	0.8	96.40%	94.01%	99.11%
0.8	0.9	96.40%	94.01%	99.11%
0.9	0.6	96.37%	93.97%	99.10%
0.9	0.7	96.37%	93.97%	99.10%
0.9	0.8	96.54%	94.28%	99.09%
0.9	0.9	96.54%	94.28%	99.09%

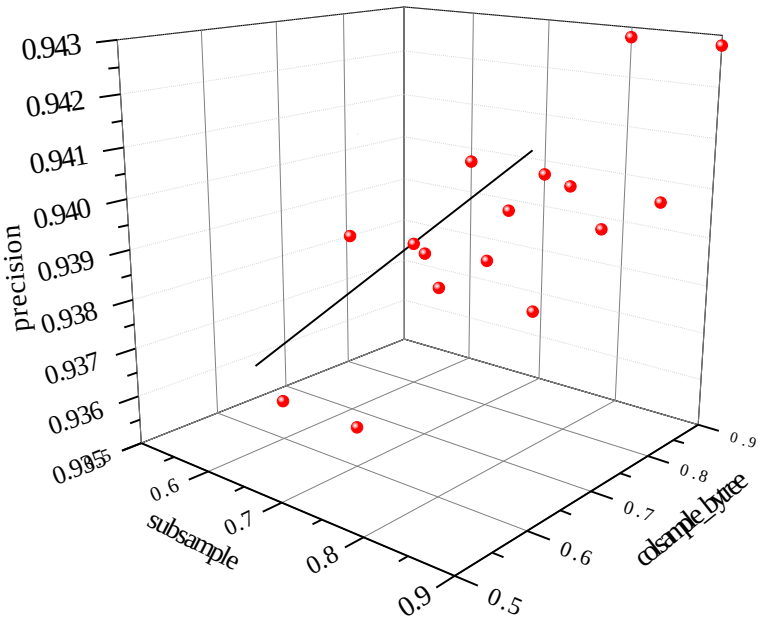

图 8. 不同 subsample 和 colsample_bytree 下模型分类精确度曲线

Fig. 8. Model precision graph for different subsample and colsample_bytree

4) reg_lambda

reg_lambda 是 XGBoost 的 L2 正则项系数。XGBoost 引入正则项降低模型复杂度，避免过拟合，提高 reg_lambda 的值使模型更保守。在实验中通过网格搜索的方法，选取最优 reg_lambda 值 5E-05（实验结果见表 8 和图 9），以达到提升模型性能的目的。

表 8. 不同 reg_lambda 下模型分类精确度表

Table 8. Model precision table for different reg_lambda

reg_lambda	accuracy	precision	recall
0.00E+00	96.47%	94.13%	99.12%
5.00E-06	96.47%	94.13%	99.12%
1.00E-05	96.47%	94.13%	99.12%
5.00E-05	96.53%	94.23%	99.12%
1.00E-04	96.51%	94.20%	99.12%
1.00E-03	96.47%	94.13%	99.12%
1.00E-01	96.51%	94.23%	99.09%
3.00E-01	96.50%	94.17%	99.12%
5.00E-01	96.43%	94.06%	99.12%
6.00E-01	96.51%	94.21%	99.12%
7.00E-01	96.50%	94.18%	99.12%
1.00E+00	96.35%	93.92%	99.11%
5.00E+00	95.20%	92.07%	98.91%
1.00E+02	71.20%	71.12%	71.34%

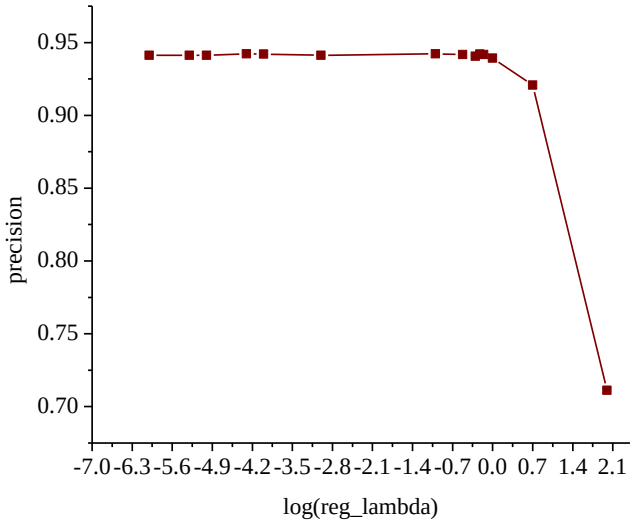

图 9. 不同 reg_lambda 下模型分类精确度曲线

Fig. 9. Model precision curve for different reg_lambda

4.4 模型性能分析

本文使用分层 K 折交叉验证的方法，使用数据集 1（70 个特征）和数据集 2（5 个特征）对模型性能进行了实验评估，使用准确率（accuracy）、精确率（precision）和召回率（recall）等指标对模型进行考量。实验结果如表 9 所示。

表 9. 模型分类效果评价表

Table 9. Model performance evaluation table

	模型	BP 神经网络		支持向量机		随机森林		XGBoost	
	数据集	1	2	1	2	1	2	1	2
调参前	Accuracy	72.15%	60.13%	N/A	N/A	98.53%	96.55%	73.37%	68.56%
	Precision	75.14%	63.16%	N/A	N/A	98.05%	94.89%	73.31%	69.37%
	Recall	66.98%	50.43%	N/A	N/A	99.03%	98.40%	73.51%	66.43%
	Time(s)	27.66	94.20	N/A	N/A	19.47	5.29	30.29	4.26
调参后	Accuracy	83.94%	60.13%	67.21%	50.08%	98.53%	96.55%	97.72%	96.67%
	Precision	82.68%	63.16%	69.34%	86.88%	98.05%	94.89%	96.47%	94.47%
	Recall	85.98%	50.43%	61.72%	20.14%	99.03%	98.40%	99.06%	99.14%
	Time(s)	109.62	94.20	29546.11	84.86	19.47	5.29	592.23	71.02

注：无 bagging 化的支持向量机使用默认超参数无法在合理时间内训练完毕，故无记录

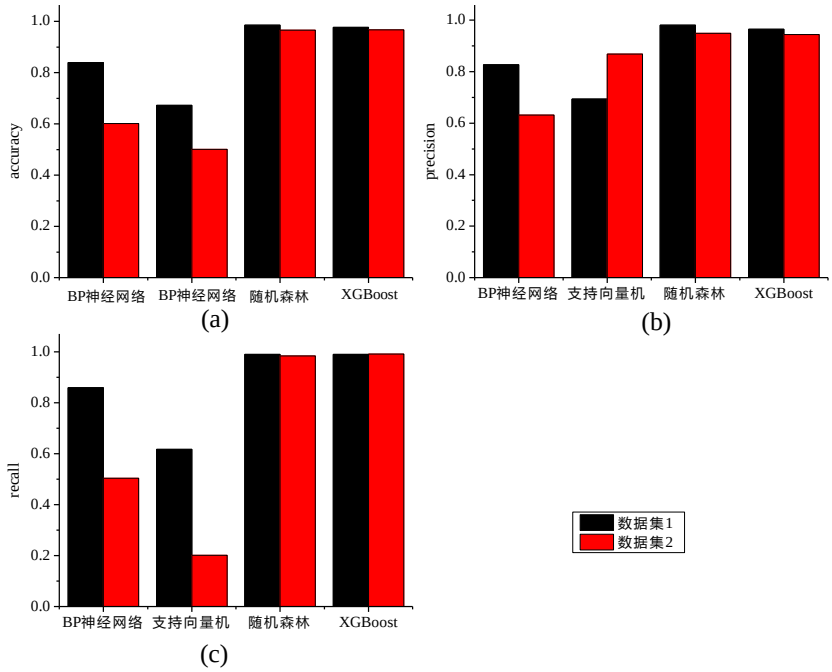

图 10. 模型分类效果评价图

Fig. 10. Model performance evaluation graph

本实验我们使用准确率、精确率和召回率等指标进行分析。在识别用户用电行为的背景下，准确率反应分类正确的比例，精确率反映判断为窃电的用户中真正窃电的用户的比例，召回率反映真正窃电的用户中被判断为窃电用户的比例。在实际

应用时，由于希望尽量少上门检查被判断为窃电用户的正常用户，即被判断为窃电用户的用户中，真正的窃电用户越多越好，精确率在此比较具有参考意义。

从表 9 可以看出随机森林算法和 XGBoost 均取得优秀的分类结果，其中随机森林算法精确率略微高于 XGBoost。随机森林优秀的分类效果和训练速度可能和其行列采样过程有关。随机森林算法在生成决策树时随机选取特征子集，这能有效降低方差，防止过拟合，同时有较快的训练速度。因此，随机森林算法通常不需要限制树深度以避免过拟合，换言之，随机森林算法对超参数的调整不敏感，默认的超参数就可以达到较好的结果。

BP 神经网络在准确率、精确率和召回率均次于随机森林和 XGBoost，且在数据集 1 上的表现明显优于数据集 2，这可能和 BP 神经网络的网络状结构相关。BP 神经网络的网络状结构被证明可以实现任何复杂非线性映射，因此适合内部机制复杂的应用场景，在数据集特征多且特征和类别（标签）关系不明晰的情况下能获得较好的训练结果。另一方面，网络结构的选择基于经验法则，缺少系统指导，使建立 BP 神经网络模型时选择合适的网络结构比较困难，这可能也是本实验中 BP 神经网络表现不佳的原因之一。

从表 9 可以看出支持向量机（SVM）在实验中训练速度慢，分类效果差。本例中使用的 LibSVM 的时间复杂度为 O(n2)，相比之下随机森林算法时间复杂度为 nlog(n)，因此 SVM 算法在数据量大的情况下训练时长大于其他算法。

结合窃电用户识别的实际应用场景，如果考虑调试和运行时间，随机森林因其速度快，精度高，超参数少的特性，比较合适；如果需要更高的分类精度，可以尝试 XGBoost，在特定数据集和正确的超参数调整后可能得到比随机森林更好的结果。

5. 总结与展望

本文对用户用电数据进行挖掘，分析并训练 BP 神经网络、支持向量机、随机森林、XGBoost 四种机器学习模型，最后选择最优算法和超参数，可判断一位用户是否有窃电嫌疑。主要研究成果如下：

（1）对主成分分析等特征选取，维数约简算法进行分析，从原始数据集处理得到两个不同数据集，应用于训练不同的机器学习算法，并分析不同数据集对不同机器学习算法训练效果的影响。

（2）研究了多种数据分类模型，分析其不同原理和实现方法对分类结果的影响。探究其最优的超参数配置，将 BP 神经网络的准确率提高到 83.94%，XGBoost 的准确率提高到 97.72%。

以下问题仍需进一步研究：

（1）本次试验中第一个数据集的特征包括三年月用电量的指标，节假日和非工作日的指标，每年前 5 个月与后 5 个月的差值的指标，两年月电量关联的指标，两年节假日电量关联的指标，两年季度电量关联的指标，用户和行业用电情况关联的指标，用电斜率和斜率差异的指标；第二个数据集使用选择日用电量平均值，日用电量最小值，日用电量最大值，日用电量方差和日用电量中位数作为特征。没有结合实际情况考虑这些变量和是否窃电之间的具体关系。在将来的研究中应结合现实中窃电行为存在的特点，尝试选取更具有意义的变量作为特征。

（2）本次试验中数据集类别不均衡，对此采取小类过抽样的方法克服，但单纯小类过抽样产生较多重复数据，提高了过拟合的可能性。在将来的研究中可尝试用 Borderline-SMOTE 与 ADASYN 等算法合成新数据[13,14]，尽量降低由于过抽样引起的过拟合的可能性。

（3）本次实验中调参过后的 BP 神经网络准确率仍然较低，SVM 训练速度较慢，分类效果不理想。将来的实验中可探究更优的超参数搜索策略，对 BP 神经网络的超参数进行进一步调整以得到更好结果；可以探究更好的维数约简算法，降低 SVM 分类器训练过程中的运算量，提高训练速度；可以结合遗传算法等其他算法，优化 BP 神经网络和 SVM 训练速度和分类效果。

6. 参考文献

[1] SALINAS S A, LI P. Privacy-preserving energy theft detection in microgrids: A state estimation approach [J]. IEEE Transactions on Power Systems, 2016, 31(2): 883-94.

[2] TARIQ M, POOR H V. Electricity Theft Detection and Localization in Grid-tied Microgrids [J]. IEEE Transactions on Smart Grid, 2016,

[3] GAUR V, GUPTA E. The determinants of electricity theft: An empirical analysis of Indian states [J]. Energy Policy, 2016, 93(127-136).

[4] 吴晓婷, 闫德勤. 数据降维方法分析与研究倡 [J]. 计算机应用研究, 2009, 26(8):

[5] PENG C, KANG Z, CHENG Q. A fast factorization-based approach to robust PCA; proceedings of the Data Mining (ICDM), 2016 IEEE 16th International Conference on, F, 2016 [C]. IEEE.

[6] MATHUR A, FOODY G M. Multiclass and binary SVM classification: Implications for training and classification users [J]. IEEE Geoscience and remote sensing letters, 2008, 5(2): 241-245.

[7] 曹峥. 反窃电系统的研究与应用 [D] [D]; 上海交通大学, 2011.

[8] LIAW A, WIENER M. Classification and regression by randomForest [J]. R news, 2002, 2(3): 18-22.

[9] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system; proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, F, 2016 [C]. ACM.

[10] GLEYZER S, MONETA L, ZAPATA O A. Development of Machine Learning Tools in ROOT; proceedings of the Journal of Physics: Conference Series, F, 2016 [C]. IOP Publishing.

[11] CHEN K, LV Q, LU Y, et al. Robust regularized extreme learning machine for regression using iteratively reweighted least squares [J]. Neurocomputing, 2017, 230(345-358).

[12] REFAEILZADEH P, TANG L, LIU H. Cross-validation [M]. Encyclopedia of database systems. Springer. 2009: 532-538.

[13] HAN H, WANG W-Y, MAO B-H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [J]. Advances in intelligent computing, 2005, 878-887.

[14] HE H, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning; proceedings of the Neural Networks, 2008 IJCNN 2008(IEEE World Congress on Computational Intelligence) IEEE International Joint Conference on, F, 2008 [C]. IEEE.

致谢

本课题由杨逸驰和徐光梓合作完成，各队员承担的工作以及贡献如下：

徐光梓撰写论文第一章（引言），对国内窃电现象的现状进行研究，结合机器学习的背景，指出窃电行为识别的背景及研究现状；撰写论文第三章（用电行为识别），研究分析多种不同现有分类算法的原理和实现过程，依据不同算法的不同实现指出支持向量机、BP 神经网络、随机森林和 XGboost 超参数的意义和影响，指导实验设计。

杨逸驰撰写论文第二章（用电行为描述），根据现有数据集指出一系列数据特征指标，描述对特征集进行维度归约的方法，解释模型评价指标的意义；撰写第四章（实验结果与分析），设计、展开、总结本次课题中的实验，首先描述现有数据集，对数据集进行预处理，然后分别训练不同分类器，结合第三章中总结的分类器原理和特点分析不同超参数对分类效果的影响，优化分类器分类效果并提出导致分类器效果差异的可能原因；撰写第五章（展望和总结），总结本课题的研究成果并提出有待研究的问题。

在本课题的开展过程中，邓水光教授和徐亦飞博士对课题研究方向的选择提出了宝贵意见，指导队员收集和查阅文献资料的方法，讲解论文书写的基本格式和技巧，在此对两位导师致以衷心的感谢。

学术诚信声明

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员： 杨逸驰 徐光梓　　指导老师：徐亦飞 邓水光

2017 年 **8** 月 **30** 日