客户异常用电行为分析

类型：最终报告

学院：计算机学院

专业：计算机科学与技术

组员/学号：吴旭辉/2120161063

王帅/2120161056

庞荣/2120161028

1. 引言

社会经济的发展使得社会用电量逐年增加，受利益驱使,窃电现象也日益严重。窃电行为不仅给供电企业造成了重大经济损失，也严重影响了正常的供用电秩序。根据国家电网公司统计，近年因窃电导致的损失达上千万元。在电力的供应与使用过程中，用电客户以减少计量电能和降低缴纳电费为目的，采用欠压法、欠流法、移相法和扩差法等手段进行电能盗窃，给供电单位造成巨大经济损失，而且还会带来一系列的用电安全隐患和社会问题。近年来，窃电方式也由野蛮窃电发展到设备智能化、手段专业化、行为隐蔽化、实施规模化的高科技窃电，给反窃电工作进一步增加了很大的难度，但反窃电手段仍以人工稽核为主,存在工作量大、取证困难和缺乏针对性等问题。随着电力系统升级，智能电力设备的普及，国家电网公司可以实时收集海量的用户用电行为数据、电力设备监测数据，因此，国家电网公司希望通过大数据分析技术，科学的开展防窃电监测分析，以提高反窃电工作效率，降低窃电行为分析的时间及成本。本文中利用数据挖掘的算法来挖掘用户用电信息，并以此来得到哪些用户存在窃电的情况从而加以监管。

2. 相关工作

现阶段反窃电的措施方法主要包括技术手段和管理手段。

1)技术手段主要是通过仪表硬件设备进行反窃电。这种技术手段一般是应用新型带防窃电作用的计量柜，当用户存在窃电行为时，通过设备装置记录失压、失流、电流不平衡、逆相序等事件，从中发现窃电的蛛丝马迹。但是改装或者升级一个地区的不同用户计量装置，这样缺乏目标性，且需要耗费大量经费和时间，在现有条件下难以全部实现。

2)管理手段主要是实行反窃电与线损考核相结合的方法，奖勤罚懒，充分调动基层电管人员的积极性，发动工作人员一起现场定量检测查找窃电点。但面对大量的用电客户，逐个巡查效率很低，最主要的缺点就是不能有效取证、准确定量、及时反馈信息。

类比中可发现，这两类传统反窃电侦查方法普遍耗时耗力，有时需要依据专业人员人工分析数据进行判断，尤其是对于一些临时性窃电，搜集证据工作十分困难。

因此，为了可以快速有效地判断出用户是否存在窃电的行为，本次实验将会尝试利用用户的用电情况基于数据挖掘的方式来判断用户是否窃电。接下来就是关于具体的算法实现的过程。

3. 算法实现

3.1 数据格式

本次项目中的数据是从以往的比赛中得到，主要的数据内容涉及每个用户的编号，以及该用户一年的用电量，具体的数据格式和类型如下：

用户清单信息

字段代码	字段名称	数据类型	是否非空
CONS_NO	用户编号	NUMBER(16)	是

CHK_STATE	窃电标识	NUMBER(10)	否

用户日用电量（记录所有用户每日用电量以及当天及前一天的电能表示值）

字段代码	字段名称	数据类型	是否非空
CONS_NO	用户编号	NUMBER(16)	是
DATA_DATE	日期	DATE	
KWH_READING	当天电能表示值	NUMBER(11,4)	
KWH_READING1	前一天电能表示值	NUMBER(11,4)	
KWH	用电量	NUMBER(11,4)	

用户编号是用于标识每一个用户的标记，窃电标识是表示该用户是否窃电，在训练集中是明确标记好了是否该用户窃电。用户日用电量是记录了每个用户一段时间的用电情况，包括时间、当天的电能表示值、前一天的电能表示值，当天的用电量=当前示值-前一天示值。训练的特征主要是从用户这一段时间的用电量中提取。

3.2 数据清洗

数据清洗，是整个数据分析过程中不可缺少的一个环节，其结果质量直接关系到模型效果和最终结论。在本次的项目中主要是对数据做值的缺省填充以及对重复数据的去重。

缺省值预处理：因为有些用户缺失某些日期的用电量 且有些数据条目里的日用电量有缺失 所以需要对数据进行补全。补全的方法主要有两个 一个是对缺失的数据全用-1 或 null 补全 第二个是利用电表读数取有记录的 最近的前后两个日期算出缺失时间段的总用电量再将总用电量平均作为缺失时间段的日用电量。例如：2016 年 7 月份的用户用电量缺失，那么就利用 8 月份的第一天的电能显示值减去 6 月份最后一天的电能显示值获得整个 7 月份的用电量，然后用该用电量除以 31 得到 7 月份平均用电量，用该平均值来填充缺失的用电量。

数据重复：在数据集中还会出现很多重复的数据，这类重复的数据也会影响到最后训练的结果，所以需要做去重的预处理。在本次项目中是将用户编号和日期共同作为主键，然后利用数据库的主键约束来去除数据集中的重复数据。

3.3 特征提取

特征提取对于数据挖掘来说非常重要。好的特征能够提升模型的性能，更能帮助我们理解数据的特点、底层结构，这对进一步改善模型、算法都有着重要作用。

特征提取主要有两个功能：

1.减少特征数量、降维，使模型泛化能力更强，减少过拟合

2.增强对特征和特征值之间的理解

3.3.1 提取的特征指标

由于直接使用用户每天的用电情况作为特征的话，得到的特征向量的维数将会非常的巨大，就会造成维度灾难，因此必需通过特征提取方法将特征维度降下来。在本次的项目中主

要是要选择四类比较有代表性的特征，分别是**周统计指标**、**月统计指标**、**季度用电量比值**和**总用电量特征**

周统计指标是指每周用电量的最大值、最小值、平均值、标准差、中位数等统计学指标，周的划分粒度比天要大一些，而且也比较可以代表一定的用户用电量信息，信息量足够丰富。同样的，月统计指标是指每月用电量的最大值、最小值、平均值、标准差、中位数等统计学指标，而且这些统计指标（最大值、最小值、平均值、标准差、中位数）也可以很好的反应用户的用电情况。

之后还对周、月用电量进行了移动平均线处理，得到了周、月相应的趋势指标。移动平均线（ Moving Average -MA）也叫移动平均价， 是利用统计学上移动平均数的原理，将过去一定天数的用电情况加以(加权)平均，连贯所得出的用电量。将这类趋势指标加入特征可以更能表现不同用户的用电情况。

第三类特征是季度用电量，季度用电量比值包括了夏秋、冬春、夏春用电量占全年用电量的比值。这主要考虑到夏天由于天气比较热，很多家庭都会用高功率的空调来纳凉，这无形之间就增加了用电量，所以全年的用电特征应该呈现夏天比较高的特征，如果出现夏天用电和其他季度相同或是比其他季度还低的话，说明该用户窃电。基于这样的考虑，所以将该统计量作为要训练的特征。

最后一类特征是总用电量特征，总用电量特征包括了用电总量、数据缺失天数、数据重复次数、以及天用电量的统计指标，这一类特征是从总体上来对数据进行特征的提取，数据缺失天数和数据重复次数可以反应一些用户是否篡改用电信息的情况，例如说数据缺失可能是用户蓄意破坏电表所导致的等等，这些数据在预处理的过程中可以获得。

最后将这四类的特征数据整合成一个高维向量。

3.3.2 特征筛选

上述将统计数据的指标作为了特征向量的组成部分，但是由于该特征维度比较大，而且有一些特征对结果的影响可能不是很大，白白增加一些计算量。所以在训练之前先要做一步特征的筛选。

在本次的项目中主要是利用带 L1 和 L2 的逻辑回归进行特征筛选，主要的原理是：利用 L1 的逻辑回归训练出来的参数中会出现 0，这也就表明 0 参数对应的特征对结果预测没有影响，同样的用 L2 的逻辑回归训练出来的参数虽然没有 0，但是会有一些特征的参数明显比别的参数下，这也就说明该参数对应的特征也对预测结果影响不大，所以这一类的特征就可以剔除出去了。

最后得到了一个具有代表性的特征向量了。

3.4 模型构建

数据挖掘目的就是要根据大量的数据构建一个可以准确预测未知情况的模型，当前训练模型的算法有很多，很多算法都可以训练出一个可以准确预测数据的模型。在本次的项目中我们利用逻辑回归、随机森林、gbdt 和 xgboost 四种训练模型的算法分别进行训练，将最后的结果按一定的加权平均得到最后的预测结果。

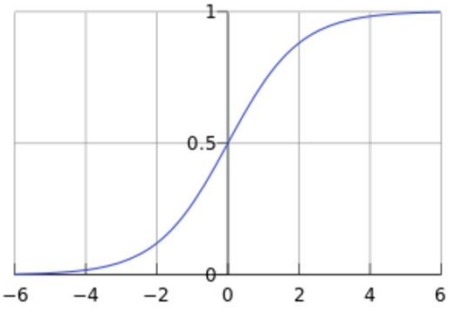
3.4.1 逻辑回归

首先是利用的逻辑回归的算法，逻辑回归(Logistic Regression, LR)又称为逻辑回归分析，是分类和预测算法中的一种。通过历史数据的表现对未来结果发生的概率进行预测。

Logistic 回归与多重线性回归实际上有很多相同之处，最大的区别就在于它们的因变量不同，其他的基本都差不多。Logistic 回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更加容易解释。所以实际中最常用的就是二分类的 Logistic 回归。Regression 问题的常规步骤为：1.寻找 h 函数（即 hypothesis）；2.构造 J 函数（损失函数）；3 想办法使得 J 函数最小并求得回归参数（θ）。

在 Logistic 回归算法中的因变量是 sigmoid 函数，如下图所示：

$$g(z)=\frac{1}{1+e^{-z}}$$



从图中可以看到 sigmoid 函数中间有一个快速变化的过程，这就可以用于做二分类的问题，也就是说当预测函数的结果高于某个阈值就为 A 类低的话就为 B 类。由此将特征向量和参数引入得到以下的预测函数。

$$\theta_0+\theta_1x_1+,\ldots,+\theta_nx_n=\sum_{i=1}^{n}\theta_ix_i=\theta^T$$

$$h_\theta(x)=g(\theta^Tx)=\frac{1}{1+e^{-\theta^Tx}}$$

最后通过线性回归的损失函数来训练最后的模型。

3.4.2 随机森林

之后是利用随机森林再次训练模型。随机森林，指的是利用多棵树对样本进行训练并预测的一种分类器。该分类器最早由 Leo Breiman 和 Adele Cutler 提出，并被注册成了商标。简单来说，随机森林就是由多棵 CART（Classification And Regression Tree）构成的。对于每棵树，它们使用的训练集是从总的训练集中有放回采样出来的，这意味着，总的训练集中的有些样本可能多次出现在一棵树的训练集中，也可能从未出现在一棵树的训练集中。在训练每棵树的节点时，使用的特征是从所有特征中按照一定比例随机地无放回的抽取的，根据 Leo Breiman 的建议，假设总的特征数量为 M，这个比例可以是 sqrt(M),1/2sqrt(M),2sqrt(M),

所以在本次的实验中 M = 85，每次采样的数量是

(1)给定训练集 S，测试集 T，特征维数 F。确定参数：使用到的 CART 的数量 t，每棵树的深度 d，每个节点使用到的特征数量 f，终止条件：节点上最少样本数 s，节点上最少的信息增益 m。

(2)从 S 中有放回的抽取大小和 S 一样的训练集 S(i)，作为根节点的样本，从根节点开始训练

(3)如果当前节点上达到终止条件，则设置当前节点为叶子节点，如果是分类问题，该叶子节点的预测输出为当前节点样本集合中数量最多的那一类 c(j)，概率 p 为 c(j)占当前样本集的比例；如果是回归问题，预测输出为当前节点样本集各个样本值的平均值。然后继续训练其他节点。如果当前节点没有达到终止条件，则从 F 维特征中无放回的随机选取 f 维特征。利用这 f 维特征，寻找分类效果最好的一维特征 k 及其阈值 th，当前节点上样本第 k 维特征小于 th 的样本被划分到左节点，其余的被划分到右节点。继续训练其他节点。

(4)重复(2)(3)直到所有节点都训练过了或者被标记为叶子节点。

(5)重复(2),(3),(4)直到所有 CART 都被训练过。

3.4.3 GBDT

GBDT 是一个应用很广泛的算法，可以用来做分类、回归。在很多的数据上都有不错的效果。GBDT 这个算法还有一些其他的名字，比如说 MART(Multiple Additive Regression Tree)，GBRT(Gradient Boost Regression Tree)，Tree Net 等，其实它们都是一个东西。

Gradient Boost 其实是一个框架，里面可以套入很多不同的算法，可以参考一下机器学习与数学(3)中的讲解。Boost 是"提升"的意思，一般 Boosting 算法都是一个迭代的过程，每一次新的训练都是为了改进上一次的结果。

原始的 Boost 算法是在算法开始的时候，为每一个样本赋上一个权重值，初始的时候，大家都是一样重要的。在每一步训练中得到的模型，会使得数据点的估计有对有错，我们就在每一步结束后，增加分错的点的权重，减少分对的点的权重，这样使得某些点如果老是被分错，那么就会被“严重关注”，也就被赋上一个很高的权重。然后等进行了 N 次迭代（由用户指定），将会得到 N 个简单的分类器（basic learner），然后我们将它们组合起来（比如说可以对它们进行加权、或者让它们进行投票等），得到一个最终的模型。

而 Gradient Boost 与传统的 Boost 的区别是，每一次的计算是为了减少上一次的残差(residual)，而为了消除残差，我们可以在残差减少的梯度(Gradient)方向上建立一个新的模型。所以说，在 Gradient Boost 中，每个新的模型的简历是为了使得之前模型的残差往梯度方向减少，与传统 Boost 对正确、错误的样本进行加权有着很大的区别。

GB 算法中最典型的基学习器是决策树，尤其是 CART，正如名字的含义，GBDT 就是 GB 和 DT 的结合。

GBDT 算法的流程如下：

Algorithm 6: $\mathbf{L}_K$-TreeBoost

$F_{k0}(\mathbf{x}) = 0, \quad k = 1, K$

For $m = 1$ to M do:

$p_k(\mathbf{x}) = \exp(F_k(\mathbf{x})) \Big/ \sum_{l=1}^{K} \exp(F_l(\mathbf{x})), \quad k = 1, K$

For $k = 1$ to K do:

$\tilde{y}_{ik} = y_{ik} - p_k(\mathbf{x}_i), \quad i = 1, N$

$\{R_{jkm}\}_{j=1}^{J} = J$-terminal node $tree(\{\tilde{y}_{ik}, \mathbf{x}_i\}_1^N)$

$\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{\mathbf{x}_i \in R_{jkm}} \tilde{y}_{ik}}{\sum_{\mathbf{x}_i \in R_{jkm}} |\tilde{y}_{ik}|(1-|\tilde{y}_{ik}|)}, \quad j = 1, J$

$F_{km}(\mathbf{x}) = F_{k,m-1}(\mathbf{x}) + \sum_{j=1}^{J} \gamma_{jkm} 1(\mathbf{x} \in R_{jkm})$

endFor

endFor

end Algorithm

0. 表示给定一个初始值。

1. 表示建立 M 棵决策树（迭代 M 次）

2. 表示对函数估计值 F(x)进行 Logistic 变换，其中 Logistic 变换的公式变换如下

$$p_k(\mathbf{x}) = \exp(F_k(\mathbf{x})) \Big/ \sum_{l=1}^{K} \exp(F_l(\mathbf{x}))$$

3. 表示对于 K 个分类进行 4-7 步的操作(其实这个 for 循环也可以理解为向量的操作，每一个样本点 xi 都对应了 K 种可能的分类 yi，所以 yi, F(xi), p(xi)都是一个 K 维的向量，这样或许容易理解一点)

4. 表示求得残差减少的梯度方向

5. 表示根据每一个样本点 x，与其残差减少的梯度方向，得到一棵由 J 个叶子节点组成的决策树

6. 为当决策树建立完成后，通过这个公式，可以得到每一个叶子节点的增益（这个增益在预测的时候用的）

7. 将当前得到的决策树与之前的那些决策树合并起来，作为新的一个模型

3.4.4 xgboost

Xgboost 是 GB 算法的高效实现，xgboost 中的基学习器除了可以是 CART（gbtree）也可以是线性分类器（gblinear）。xgboost 算法的步骤和 GB、GBDT 基本相同，都是首先初始化为一个常数，gb 是根据一阶导数 ri，xgboost 是根据一阶导数 gi 和二阶导数 hi，迭代生成基学习器，相加更新学习器。但是与 GB 算法相比它做了许多的优化，其中包括：

(1). xgboost 在目标函数中显示的加上了正则化项，基学习为 CART 时，正则化项与树的叶子节点的数量 T 和叶子节点的值有关。

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

$$\text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

(2). GB 中使用 Loss Function 对 f(x)的一阶导数计算出伪残差用于学习生成 fm(x)，xgboost 不仅使用到了一阶导数，还使用二阶导数。

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

第 t 次的 loss：

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

对上式做二阶泰勒展开：g 为一阶导数，h 为二阶导数

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

$$\text{where } g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \text{ and } h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

(3). 上面提到 CART 回归树中寻找最佳分割点的衡量标准是最小化均方差，xgboost 寻找分割点的标准是最大化，lamda，gama 与正则化项相关

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma$$

3.4.5 参数选择

在这次实验中为了可以获得准确的训练参数，我们使用了网格搜索（GridSearch）来调整参数。

模型参数选择的范围如下：

```
n_components = [15, 20, 25, X.shape[1]]
Cs = logspace(-4, 2, 50)

max_depth_rf = [11, 13, 15, 17]
n_estimators_rf = [1100, 1500, 2500, 3000, 3500]

max_depth_gbdt = [3, 5, 7, 9]
learning_rate_gbdt = [0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
n_estimators_gbdt = [130, 150, 170, 190, 210, 230, 250]

max_depth_xgb = [7, 9, 11, 13, 15, 17]
n_estimators_xgb = [150, 70, 210, 250, 290]
learning_rate_xgb = [0.05, 0,1, 0,15, 0.2]
gamma_xgb = [0.1, 0.2]
min_child_weight_xgb = [1, 3, 5, 7]
max_delta_step_xgb = [0.5, 1, 1.5]
subsample_xgb = [0.5, 0.7, 0.9, 1]
colsample_bytree_xgb = [0.5, 0.7, 0.9, 1]
reg_lambda_xgb = [1, 2, 50, 100, 300]
```

通过网格搜索后训练得到了最优的模型，最后计算四种模型预测结果的平均值来得到最终的结果（实验中发现四种模型用相同的权重效果会好一些）

3.4.6 评估标准

本次实验中使用的评估标准是 map，MAP:全称 mean average precision(平均准确率)。其计算的方式与该系统的准确率和召回率是相关的，公式如下：

$$mAP = \int_0^1 P(R) dR$$

其中 R 表示召回率，P(R)表示 R 对应的准确率，通过上述的积分就计算出了平均准确率，mAP 是为解决 P，R，F-measure 的单点值局限性的，同时考虑了检索效果的排名情况。

4.实验结果

算法实验的过程中将数据集按 6：2：2 分别作为训练集，验证集和测试集
测试的结果为：

指标	得分
MAP score	0.729860045624
ROC score	0.931859479104
Precision rate	0.678445229682
Recall rate	0.755905511811

参考文献

[1]. 翟莺鸽．反窃电风险控制方法研究[D]．上海：上海交通大学，2014：36-52．ZHAI Y G．Study on anti-stealing electric power control method[D]．Shanghai：Shanghai Jiao Tong University，2014：36-52．

[2]. HASHMI M U，PRIOLKAR J G．Anti-theft energy metering for smart electrical distribution system[D]．2015 International Conference on Industrial Instrumentation and Control(ICIC)，2015：1424-1428．

[3]. MEFFE A，de OLIVEIRACCB．Technical loss calculation by distribution system segment with corrections from measurements[c]．20th International Conference and Exhibition on Electricity Distribution，2009：1-4．

[4]. LV Z．Design and development of all intelligent anti—steal—electricity—power instrument[J]．Electronic Measurement Technology，2006，29(5)：93-95．

[5]. LI M .Discussion on Anti—Stealing Electric Power of Power Supply Enterprise[M] .Beijing：China Academic Journal Electronic Publishing House，2008：291-292．

[6]. FANG W．Recommendations for how to strengthen the electricity check the anti·stealing electric power work[J]．Guangdong Science and Technology，2012(19)：78-82．