第6章 电力窃漏电用户自动识别

6.1背景与挖掘目标

传统的防窃漏电方法主要通过定期巡检、定期校验电表、用户举报窃电等手段来发现窃电或计量装置故障。但这种方法对人的依赖性太强，抓窃查漏的目标不明确。目前很多供电局主要通过营销稽查人员、用电检查人员和计量工作人员利用计量异常报警功能和电能量数据查询功能开展用户用电情况的在线监控工作，通过采集电量异常、负荷异常、终端报警、主站报警、线损异常等信息，建立数据分析模型，来实时监测窃漏电情况和发现计量装置的故障。根据报警事件发生前后客户计量点有关的电流、电压、负荷数据情况等，构建基于指标加权的用电异常分析模型，实现检查客户是否存在窃电、违章用电及计量装置故障等。

以上防窃漏电的诊断方法，虽然能获得用电异常的某些信息，但由于终端误报或漏报过多，无法达到真正快速精确定位窃漏电嫌疑用户的目的，往往令稽查工作人员无所适从。而且在采用这种方法建模时，模型各输入指标权重的确定需要用专家的知识和经验，具有很大的主观性，存在明显的缺陷，所以实施效果往往不尽如人意。

现有的电力计量自动化系统能够采集到各相电流、电压、功率因数等用电负荷数据以及用电异常等终端报警信息。异常告警信息和用电负荷数据能够反映用户的用电情况，同时稽查工作人员也会通过在线稽查系统和现场稽查来查找出窃漏电用户，并录入系统。若能通过这些数据信息提取出窃漏电用户的关键特征，构建窃漏电用户的识别模型，就能自动检查判断用户是否存在窃漏电行为。

表 6-1 给出了某企业大用户的用电负荷数据，采集时间间隔为 15 分钟，即 0.25 小时，可进一步计算该大用户的用电量，

表 6-2 给出了该企业大用户的终端报警数据，其中与窃漏电相关的报警能较好的识别用户的窃漏电行为，表 6-3 给出了某企业大用户违约、窃电处理通知书，里面记录了用户的用电类别和窃电时间。

表 6-1 某企业大用户用电负荷数据

用户编号	时间	有功总	B相	C相	电流A相	电流B相	电流C相	电压A相	电压B相	电压C相	功率因数	功率因数A	功率因数B	功率因数C
0319001000019011001	2011/11/10	202	0	349.2	33.6	0	33.4	10500	0	10500	0.784	0.573	-10000	0.996
0319001000019011001	2011/11/10 0:15	194.8	0	355.4	32.4	0	34	10500	0	10500	0.789	0.573	-10000	0.996
0319001000019011001	2011/11/10 0:30	210.4	0	366	35	0	35	10500	0	10500	0.784	0.573	-10000	0.996
0319001000019011001	2011/11/10 0:45	199.6	0	376.4	33.2	0	36	10500	0	10500	0.793	0.573	-10000	0.996
0319001000019011001	2011/11/10 1:00	191.2	0	334.6	31.8	0	32	10500	0	10500	0.785	0.573	-10000	0.996
0319001000019011001	2011/11/10 1:15	192.4	0	340.8	32	0	32.6	10500	0	10500	0.786	0.573	-10000	0.996
0319001000019011001	2011/11/10 1:30	192.4	0	353.4	32	0	33.8	10500	0	10500	0.79	0.573	-10000	0.996
0319001000019011001	2011/11/10 1:45	197.2	0	357.6	32.8	0	34.2	10500	0	10500	0.789	0.573	-10000	0.996
0319001000019011001	2011/11/10 2:00	178	0	320.8	29.6	0	30.4	10500	0	10600	0.788	0.573	-10000	0.996
0319001000019011001	2011/11/10 2:15	173.2	0	311.6	28.8	0	29.8	10500	0	10500	0.788	0.573	-10000	0.996
0319001000019011001	2011/11/10 2:30	185.2	0	332.4	30.8	0	31.8	10500	0	10500	0.787	0.573	-10000	0.996
0319001000019011001	2011/11/10 2:45	175.6	0	326.2	29.2	0	31.2	10500	0	10500	0.791	0.573	-10000	0.996
0319001000019011001	2011/11/10 3:00	164.8	0	311.6	27.4	0	29.8	10500	0	10500	0.793	0.573	-10000	0.996
0319001000019011001	2011/11/10 3:15	185.8	0	317.8	31.2	0	30.4	10400	0	10500	0.782	0.573	-10000	0.996
0319001000019011001	2011/11/10 3:30	169.6	0	303.2	28.2	0	29	10500	0	10500	0.787	0.573	-10000	0.996
0319001000019011001	2011/11/10 3:45	179.2	0	320	29.8	0	30.6	10500	0	10500	0.787	0.573	-10000	0.996
0319001000019011001	2011/11/10 4:00	175.6	0	305.2	29.2	0	29.2	10500	0	10500	0.784	0.573	-10000	0.995
0319001000019011001	2011/11/10 4:15	178.6	0	324	30	0	31	10400	0	10500	0.788	0.572	-10000	0.995
0319001000019011001	2011/11/10 4:30	173.2	0	313.6	28.8	0	30	10500	0	10500	0.788	0.573	-10000	0.996
0319001000019011001	2011/11/10 4:45	166	0	297	27.6	0	28.4	10500	0	10500	0.787	0.573	-10000	0.996
0319001000019011001	2011/11/10 5:00	170.8	0	303.2	28.4	0	29	10500	0	10500	0.786	0.573	-10000	0.996
0319001000019011001	2011/11/10 5:15	176.8	0	322	29.4	0	30.8	10500	0	10500	0.789	0.573	-10000	0.996
0319001000019011001	2011/11/10 5:30	175.6	0	301	29.2	0	28.8	10500	0	10500	0.783	0.573	-10000	0.995
0319001000019011001	2011/11/10 5:45	164.4	0	299	27.6	0	28.6	10400	0	10500	0.789	0.573	-10000	0.996
0319001000019011001	2011/11/10 6:00	168.4	0	315.8	28	0	30.2	10500	0	10500	0.792	0.573	-10000	0.996
0319001000019011001	2011/11/10 6:15	165.6	0	284.4	27.8	0	27.2	10400	0	10500	0.783	0.573	-10000	0.996
0319001000019011001	2011/11/10 6:30	164.4	0	297	27.6	0	28.4	10400	0	10500	0.788	0.573	-10000	0.996
0319001000019011001	2011/11/10 6:45	188.2	0	334.6	31.6	0	32	10400	0	10500	0.787	0.573	-10000	0.996
0319001000019011001	2011/11/10 7:00	179.8	0	315.8	30.2	0	30.2	10400	0	10500	0.785	0.572	-10000	0.996
0319001000019011001	2011/11/10 7:15	165.6	0	290	27.8	0	28	10400	0	10400	0.785	0.573	-10000	0.996
0319001000019011001	2011/11/10 7:30	219	0	391	36.4	0	37.4	10500	0	10500	0.787	0.573	-10000	0.996
0319001000019011001	2011/11/10 7:45	227.6	0	403.6	38.2	0	38.6	10400	0	10500	0.786	0.573	-10000	0.996

表 6-2 某企业大用户终端报警信息

用户名称	时间	计量点 ID	报警编号	报警名称
某企业大用户	2010/4/1 0:01	0319001000045110001	135	最大需量复零
某企业大用户	2010/4/2 18:44	0319001000045110001	152	电流不平衡
某企业大用户	2010/4/2 18:47	0319001000045110001	143	A 相电流过负荷
某企业大用户	2010/4/2 18:47	0319001000045110001	145	C 相电流过负荷
某企业大用户	2010/4/2 21:07	0319001000045110001	152	电流不平衡
某企业大用户	2010/4/2 21:22	0319001000045110001	145	C 相电流过负荷
某企业大用户	2010/4/2 21:25	0319001000045110001	143	A 相电流过负荷
某企业大用户	2010/4/3 5:45	0319001000045110001	145	C 相电流过负荷

*由于各方面原因，终端报警存在一定误报和漏报情况。

表 6-3 某企业大用户违约、窃电处理通知书

<table>
<tr><td rowspan="3">用户基本信息</td><td>用户名称</td><td colspan="2">某企业大用户</td><td>用户编号</td><td colspan="3">7210100429</td></tr>
<tr><td>用电地址</td><td colspan="2">******</td><td>用电类别</td><td>大工业</td><td>报装容量</td><td>1515</td></tr>
<tr><td>计量方式</td><td>高供高计</td><td>电流互感器变比</td><td>100/5</td><td>电压互感器变比</td><td colspan="2">10kV/100V</td></tr>
<tr><td>现场情况</td><td colspan="7">我局用电检查人员根据群众举报，于 2014 年 11 月 17 日到你户进行用电检查，发现你户（客户编号：7210100429）配电变压器（3 台容量为 400KVA 和 1 台容量为 315KVA）的高压计量柜的前门封印（SJL00014930）被人为破坏，计费电能表（NO：01026660；条形码 NO： SFF5104000864）的计量接线盒 C 相电压连接片被人为断开，计费电能表显示 C 相电流为 0，现场检测计费电能表 C 相同时失压失流，导致少计电量。即时报当地公安机关并拍照取证，现场对你户作停电处理。当时计费电能表抄见有功止码为 16448.77。</td></tr>
<tr><td>违约、窃电行为</td><td colspan="7">故意使供电企业用电计量装置不准或失效。</td></tr>
<tr><td>计算方</td><td colspan="7">确定依据：计量自动化系统记录（2014 年 11 月 12 日计费电能表存在失压失流记录，直至 2014 年 11 月 17 日止 C 相电压和电流数值均为 0）。
结论：现确定你户窃电时间由 2014 年 11 月 12 日起至 2014 年 11 月 17 日止，共 6 天。
根据现场计量装置检查情况，计费电能表 C 相失压失流，依据计量自动化系统召</td></tr>
</table>

法及依据	测数据分析，你户计费电能表（NO：01026660；条形码 NO： SFF5104000864）的2014-11-12 功率因数：COS(30 °+ φ)=0.572，即 φ=25.11 °,COS φ=0.905。更正系数=P 正确 / P 错误=UICOS φ/〔UICOS(φ+30 °)〕=1.732×0.905/0.572=2.74，更正率=更正系数-1=2.74-1=1.74。2014 年 11 月 12 日计费电能表记录有功止码为 16431.45，查处现场计费电能表抄见有功止码为 16448.77，电流互感器变比为 100/5,电压互感器变比为 10000/100。根据《供电营业规则》第一百零二条规定，窃电者应按所窃电量补交电费，并承担补交电费三倍的违约使用电费。具体计算如下： 1、计费电能表已计收电量=（16448.77－16431.45）×100/5×10000/100=34640（KW.h） 2、窃电电量=已计收电量×更正率=34640×1.74=60274（KW.h） 3、窃电电费=60274×0.6709=40437.82（元） 4、城市建设附加费=60274×0.014=843.84（元） 5、违约使用电费=40437.82×3=121313.46（元） 6、合计金额=40437.82+843.84+121313.46=162595.12（元）	
	合计电费： 162595.12 元	大写金额： 拾陆万贰仟伍佰玖拾伍圆壹角贰分

本次数据挖掘建模目标如下：

1、 归纳出窃漏电用户的关键特征，构建窃漏电用户的识别模型；

2、 利用实时监测数据，调用窃漏电用户识别模型实现实时诊断。

6.2分析方法与过程

窃漏电用户在电力计量自动化系统的监控大用户中只占小部分，同时某些大用户也不可能存在窃漏电行为，如银行、税务、学校、工商等非居民类别，故在数据预处理时有必要将这些类别用户剔除。系统中的用电负荷不能直接体现出用户的窃漏电行为，终端报警存在很多误报和漏报的情况，故需要进行数据探索和预处理，总结窃漏电用户的行为规律，再从数据中提炼出描述窃漏电用户的特征指标。最后结合历史窃漏电用户信息，整理出识别模型的专家样本数据集，再进一步构建分类模型，实现窃漏电用户的自动识别。

窃漏电用户识别流程如图 6-1 所示，主要包括以下步骤：

1） 从电力计量自动化系统、营销系统有选择性地抽取部分大用户用电负荷、终端报警及违约窃电处罚信息等原始数据；

2） 对样本数据探索分析，剔除不可能存在窃漏电行为行业的用户，即白名单用户，初步审视正常用户和窃漏电用户的用电特征；

3） 对样本数据进行预处理，包括数据清洗、缺失值处理和数据变换；

4） 构建专家样本集；

5） 构建窃漏电用户识别模型；

6） 在线监测用户用电负荷及终端报警，调用模型实现实时诊断。

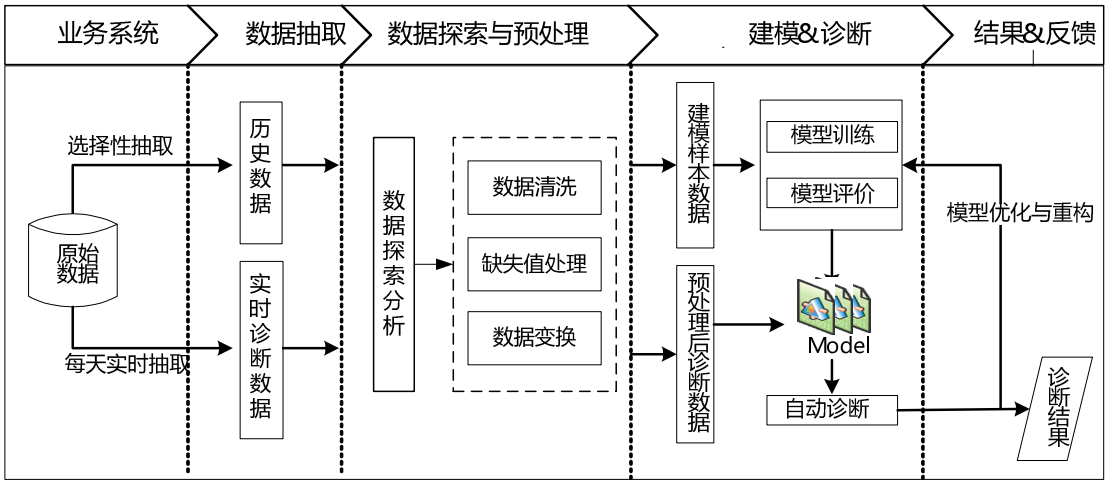

图 6-1 窃漏电用户识别流程

6.2.1 数据抽取

与窃漏电相关的原始数据主要有用电负荷数据、终端报警数据、违约窃电处罚信息以及用户档案资料等，故进行窃漏电诊断建模时需从营销系统和计量自动化系统中抽取如下数据：

1、从营销系统抽取的数据主要有：

- 用户基本信息：用户名称、用户编号、用电地址、用电类别、报装容量、计量方式、电流互感器变比、电压互感器变比；
- 违约、窃电处理记录；
- 计量方法及依据。

2、从计量自动化系统采集的数据属性主要有：

- 实时负荷：时间点、计量点、总有功功率、A/B/C 相有功功率、A/B/C 相电流、A/B/C 相电压、A/B/C 相功率因数；
- 终端报警。

为了尽可能全面覆盖各种窃漏电方式，建模样本要包含不同用电类别的所有窃漏电用户及部分正常用户。窃漏电用户的窃漏电开始时间和结束时间是表征其窃漏电的关键时间节点，在这些时间节点上，用电负荷和终端报警等数据也会有一定的特征变化，故样本数据抽取时务必要包含关键时间节点前后一定范围的数据，并通过用户的负荷数据计算出当天的用

电量，公式如下：

$$f_l = 0.25 \sum_{m_i \in l天} m_i \tag{6-1}$$

其中 f_l 为第 l 天的用电量，m_i 为第 l 天每隔 15 分钟的总有功功率，对其累加求和得到当天用电量。

基于此，本案例抽取某市近 5 年来所有的窃漏电用户有关数据和不同用电类别正常用电用户共 208 个用户的有关数据，时间为 2009 年 1 月 1 日至 2014 年 12 月 31 日，同时包含每天是否有窃漏电情况的标识。

6.2.2 数据探索分析

数据探索分析是对数据进行初步研究，发现数据的内在规律特征，有助于选择合适的数据预处理和数据分析技术。本案例主要采用分布分析和周期性分析等方法对电量数据进行数据探索分析。

1. 分布分析

对 2009 年 1 月 1 日至 2014 年 12 月 31 日共 5 年所有的窃漏用户进行分布分析，统计出各个用电类别的窃漏电用户分布情况，从图 6-2 可以发现非居民类别不存在窃漏电情况，故在接下的分析不考虑非居民类别的用电数据。

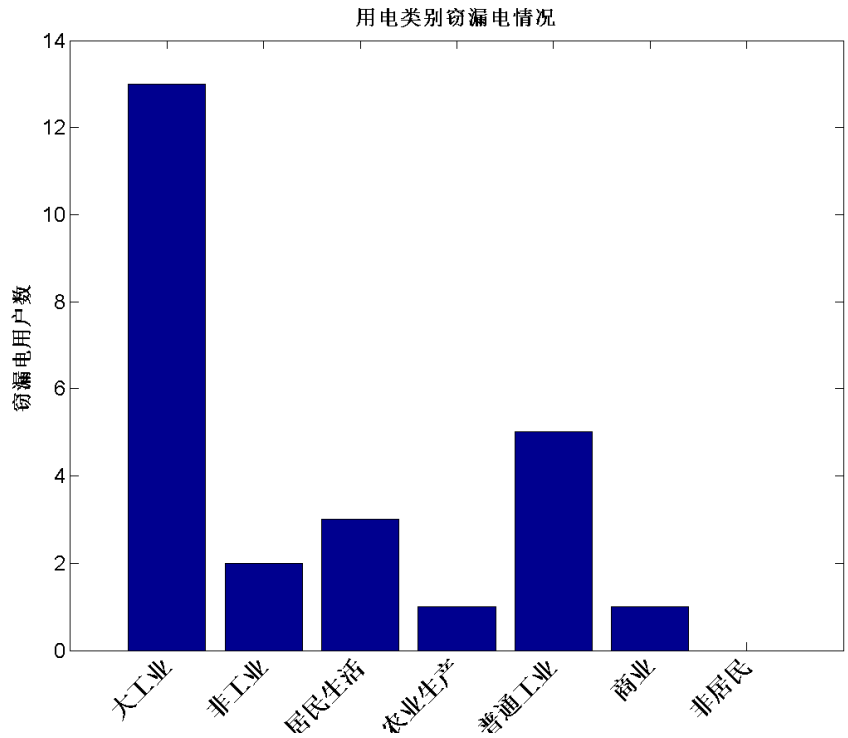

图 6-2 用电类别窃漏电情况图

2. 周期性分析

随机抽取一个正常用电用户和一个窃漏电用户，采用周期性分析对用电量进行探索。

1) 正常用电用户电量探索分析

正常用电量特征表现见图 6-3 和表 6-4。总体来看该用户用量比较平稳，没有太大的波动，这就是用户正常用电的电量指标特征。

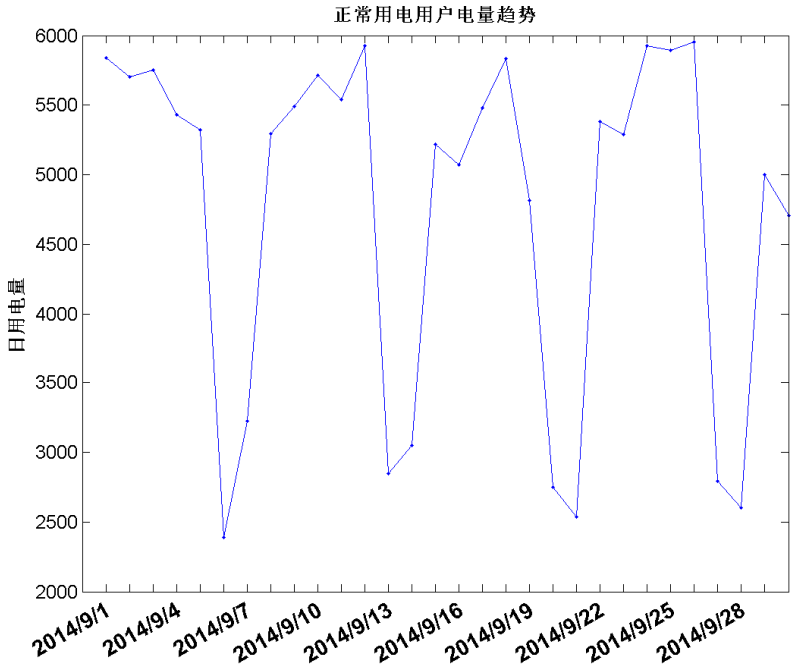

图 6-3 正常用电用户电量趋势图

表 6-4 正常用电电量数据

日期	日电量(kW)	日期	日电量(kW)
2014/9/1	5840	2014/9/16	5072
2014/9/2	5704	2014/9/17	5480
2014/9/3	5754	2014/9/18	5832
2014/9/4	5431	2014/9/19	4816
2014/9/5	5322	2014/9/20	2748
2014/9/6	2392	2014/9/21	2536
2014/9/7	3225	2014/9/22	5384
2014/9/8	5296	2014/9/23	5288
2014/9/9	5488	2014/9/24	5928
2014/9/10	5713	2014/9/25	5896
2014/9/11	5542	2014/9/26	5952
2014/9/12	5928	2014/9/27	2792
2014/9/13	2848	2014/9/28	2600
2014/9/14	3048	2014/9/29	5000
2014/9/15	5216	2014/9/30	4704

2) 窃电用电电量探索分析

窃漏电用电量特征表现见图 6-4 和表 6-5。这里可以明显看出该用户用电量出现明显下降的趋势，这就是用户异常用电的电量指标特征。

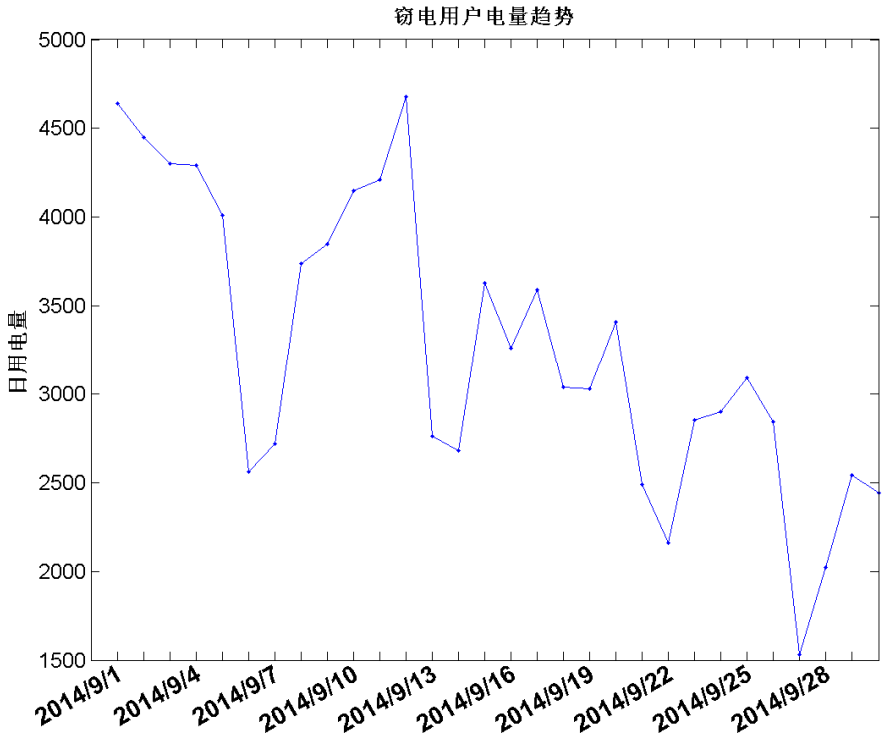

图 6-4 窃电用户电量趋势图

表 6-5 窃电用电电量数据

日期	日电量(kW)	日期	日电量(kW)
2014/9/1	4640	2014/9/16	3260
2014/9/2	4450	2014/9/17	3590
2014/9/3	4300	2014/9/18	3040
2014/9/4	4290	2014/9/19	3030
2014/9/5	4010	2014/9/20	3410
2014/9/6	2560	2014/9/21	2490
2014/9/7	2720	2014/9/22	2160
2014/9/8	3740	2014/9/23	2850
2014/9/9	3850	2014/9/24	2900
2014/9/10	4150	2014/9/25	3090
2014/9/11	4210	2014/9/26	2840
2014/9/12	4680	2014/9/27	1530
2014/9/13	2760	2014/9/28	2020
2014/9/14	2680	2014/9/29	2540
2014/9/15	3630	2014/9/30	2440

分析结论：从图 6-4 看出正常用电到窃电过程是用电量持续下降的过程，该用户从 2014 年 9 月 1 开始用电量下降，并且持续下降，这就是用户开始窃电时所表现出来的重要特征。

6.2.3 数据预处理

本案例主要从数据清洗、缺失值处理、数据变换等方面对数据进行预处理。

1. 数据清洗

数据清洗的主要目的是从业务以及建模的相关需要方面考虑，筛选出需要的数据。由于原始数据中并不是所有的数据都需要进行分析，因此需要在数据处理时，可以将赘余的数据进行过滤。本案例主要进行如下操作：

1、通过数据的探索分析，发现在用电类别中，非居民用电类别不可能存在漏电窃电的现象，需要将非居民用电类别的用电数据过滤掉。

2、结合本案例的业务，节假日用电量与工作日相比，会明显偏低。为了尽可能达到较好数据效果，过滤节假日的用电数据。

2. 缺失值处理

在原始计量数据，特别是用户电量抽取过程中，发现存在缺失的现象。若将这些值抛弃掉，会严重影响供出电量的计算结果，最终导致日线损率数据误差很大。为了达到较好的建模效果，需要对缺失值处理。本案例采用拉格朗日插值法对缺失值进行插补。

选取数据中部分数据做为实例，如表 6-6 是三个用户一个月工作日的电量数据，对缺失值采用拉格朗日插值法进行插补。

表 6-6 三个用户一个月工作日用电量数据

日期 \ 用电量	用户 A	用户 B	用户 C
2014/9/1	235.8333	324.0343	478.3231
2014/9/2	236.2708	325.6379	515.4564
2014/9/3	238.0521	328.0897	517.0909
2014/9/4	235.9063		514.89
2014/9/5	236.7604	268.8324	
2014/9/8		404.048	486.0912
2014/9/9	237.4167	391.2652	516.233
2014/9/10	238.6563	380.8241	
2014/9/11	237.6042	388.023	435.3508
2014/9/12	238.0313	206.4349	487.675
2014/9/15	235.0729		
2014/9/16	235.5313	400.0787	660.2347
2014/9/17		411.2069	621.2346
2014/9/18	234.4688	395.2343	611.3408
2014/9/19	235.5	344.8221	643.0863
2014/9/22	235.6354	385.6432	642.3482

2014/9/23	234.5521	401.6234	
2014/9/24	236	409.6489	602.9347
2014/9/25	235.2396	416.8795	589.3457
2014/9/26	235.4896		556.3452
2014/9/29	236.9688		538.347

***数据详见：/示例程序/data/missing_data.xls**

拉格朗日插值法补值，具体方法如下：

首先从原始数据集中确定因变量和自变量，取出缺失值前后 5 个数据（前后数据不足 5 个的，将仅有的数据组成一组），根据取出来的 10 个数据组成一组。然后采用拉格朗日多项式插值公式

$$L_n(x)=\sum_{i=0}^{n} l_i(x)y_i \tag{6-2}$$

$$l_i(x)=\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j} \tag{6-3}$$

其中 x 为缺失值对应的下标序号，$L_n(x)$ 为缺失值的插值结果，x_i 为非缺失值 y_i 的下标序号。对全部缺失数据依次进行插补，直到不存在缺失值为止。数据插补代码如代码清单 6-1 所示。

代码清单 6-1 拉格朗日插值代码

```
%% 拉格朗日插值算法
clear;
% 参数初始化
inputfile='../data/missing_data.xls';      % 输入数据路径,需要使用 Excel 格式；
outputfile='../tmp/missing_data_processed.xls';                    %输出数据路径,需要使用 Excel 格式

%% 拉格朗日插值
% 读入文件
data=xlsread(inputfile);
[rows,cols]=size(data);

% 按照每列进行插值处理
% 其中 ployinterp_column 为自定义函数，针对列向量进行插值
for j=1:cols
    data(:,j)=ployinterp_column(data(:,j));
end

%% 写入文件
```

```
xlswrite(outputfile,data);
```

根据代码清单 6-1 补全的数据如表 6-7 所示，斜体加粗表示补全的数据。

表 6-7 用户电量补全数据

用电量 / 日期	用户 A	用户 B	用户 C
2014/9/4	235.9063	***203.4621***	514.89
2014/9/5	236.7604	268.8324	***465.2697***
2014/9/8	***237.1512***	404.048	486.0912
2014/9/10	238.6563	380.8241	***516.233***
2014/9/15	235.0729	***237.3481***	***608.5369***
2014/9/17	***235.315***	411.2069	621.2346
2014/9/23	234.5521	401.6234	***618.1972***
2014/9/26	235.4896	***420.7486***	556.3452
2014/9/29	236.9688	***408.9632***	538.347

3. 数据变换

通过电力计量系统采集的电量、负荷虽然在一定程度上能反映用户窃漏电行为的某些规律，但要作为构建模型的专家样本，特征不明显，需要进行重新构造。基于数据变换，得到新的评价指标来表征窃漏电行为所具有的规律，其评价指标体系详见图 6-5。

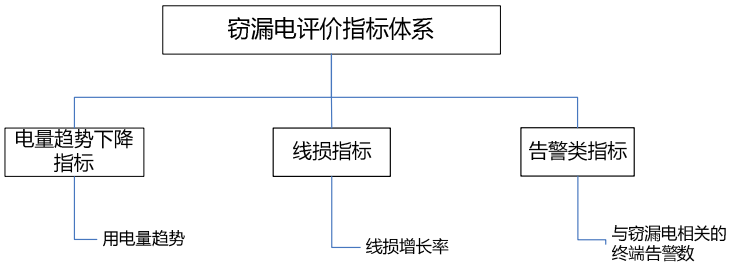

图 6-5 窃漏电评价指标体系

窃漏电评价指标如下：

1、 电量趋势下降指标

由 6.2.2 节的周期性分析可以发现，正常用户的用电量较为平稳，窃漏电用户的用电量呈现下降的趋势，然后趋于平缓，针对此可考虑前后几天作为统计窗口期，考虑期间的下降趋势，利用电量做直线拟合得到的斜率作为衡量，如果斜率随时间不断下降，那该用户的窃漏电可能性就很大，如图 6-6 所示，第一幅图展示了每天的用电量，其他图表示了随着时

间推移在各自统计窗口期以用电量做直线拟合的斜率，可以看出斜率随着时间逐步下降。

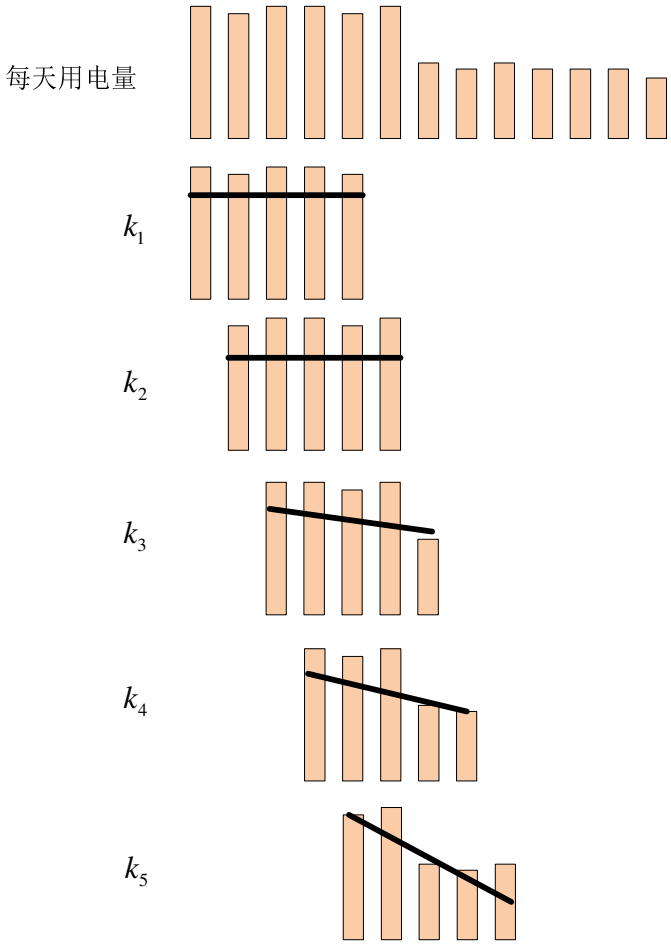

图 6-6 电量趋势下降示意图

对统计当天设定前后 5 天为统计窗口期，计算这 11 天内的电量趋势下降情况，首先计算这 11 天中每天的电量趋势，其中第i天的用电量趋势是考虑前后 5 天期间的用电量斜率，即

$$k_i = \frac{\sum_{l=i-5}^{i+5}(f_l - \overline{f})(l - \bar{l})}{\sum_{l=i-5}^{i+5}(l - \bar{l})^2} \tag{6-4}$$

其中$\overline{f} = \frac{1}{11}\sum_{l=i-5}^{i+5} f_l$，$\bar{l} = \frac{1}{11}\sum_{l=i-5}^{i+5} l$，$k_i$为第$i$天的电量趋势，$f_l$为第$l$天的用电量。

若电量趋势为不断下降的，则认为具有一定的窃电嫌疑，故计算这 11 天内，当天比前一天用电量趋势为递减的天数，即设有

$$D(i) = \begin{cases} 1, k_i < k_{i-1} \\ 0, k_i \geq k_{i-1} \end{cases} \tag{6-5}$$

则这 11 天内的电量趋势下降指标为

$$T=\sum_{n=i-4}^{i+5}D(n) \tag{6-6}$$

2、 线损指标

线损率是用于衡量供电线路的损失比例，同时可结合线户拓扑关系（如图 6-7）计算出用户所属线路在当天的线损率，一条线路上同时供给多个用户，若第l天的线路供电量为s_l，线路上各个用户的总用电量为$\sum_m f_l^{(m)}$，线路的线损率公式为

$$t_l=\frac{s_l-\sum_m f_l^{(m)}}{s_l}\times 100\% \tag{6-7}$$

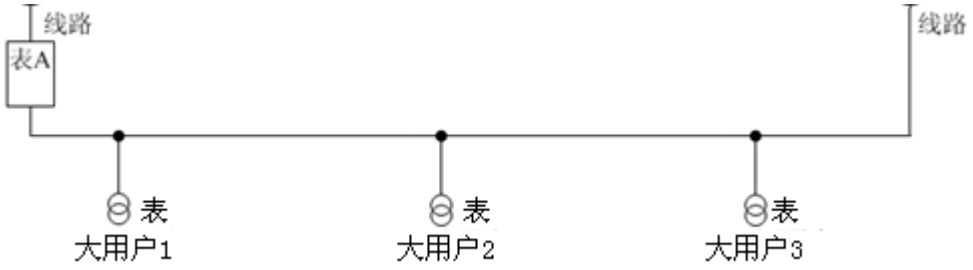

图 6-7 线路与大用户的拓扑关系示意图

线路的线损率可作为用户线损率的参考值，若用户发生窃漏电，则当天的线损率会上升，但由于用户每天的用电量存在波动，单纯以当天线损率上升了作为窃漏电特征则误差过大，所以考虑前后几天的线损率平均值，判断其增长率是否大于 1%，若线损率的增长率大于 1%则具有窃漏电的可能性。

对统计当天设定前后 5 天为统计窗口期，首先分别计算统计当天与前 5 天之间的线损率平均值V_i^1和统计当天与后 5 天之间的线损率平均值V_i^2，若V_i^1比V_i^2的增长率大于 1%，则认为具有一定的窃电嫌疑，故定义线损指标

$$E(i)=\begin{cases}1, \dfrac{V_i^1-V_i^2}{V_i^2}>1\% \\ 0, \dfrac{V_i^1-V_i^2}{V_i^2}\le 1\%\end{cases} \tag{6-8}$$

3、 告警类指标

与窃漏电相关的终端报警主要有电压缺相、电压断相、电流反极性等告警，计算发生与窃漏电相关的终端报警的次数总和，作为告警类指标。

6.2.4 构建专家样本

对 2009 年 1 月 1 日至 2014 年 12 月 31 日所有窃漏电用户及正常用户的电量、告警及线损数据和该用户在当天是否窃漏电的标识，按窃漏电评价指标进行处理并选取其中 291 个样本数据，得到专家样本库，见表 6-8。

表 6-8 专家样本数据

时间	用户编号	电量趋势下降指标	线损指标	告警类指标	是否窃漏电
2014 年 9 月 6 日	9900667154	4	1	1	1
2014 年 9 月 20 日	9900639431	4	0	4	1
2014 年 9 月 17 日	9900585516	2	1	1	1
2014 年 9 月 14 日	9900531154	9	0	0	0
2014 年 9 月 17 日	9900491050	3	1	0	0
2014 年 9 月 13 日	9900461501	2	0	0	0
2014 年 9 月 22 日	9900412593	5	0	2	1
2014 年 9 月 20 日	9900366180	3	1	3	1
2014 年 9 月 19 日	9900322960	3	0	0	0
2014 年 9 月 9 日	9900254673	4	1	0	0
2014 年 9 月 18 日	9900196505	10	1	2	1
2014 年 9 月 16 日	9900145248	10	1	3	1
2014 年 9 月 6 日	9900137535	2	0	3	0
2014 年 9 月 7 日	9900064537	4	0	2	0
2014 年 9 月 9 日	9110103867	3	0	0	0
2014 年 9 月 23 日	9010100689	0	0	3	0
2014 年 9 月 21 日	8910101840	9	0	3	1
2014 年 9 月 11 日	8910101209	0	0	2	0
2014 年 9 月 19 日	8910101132	8	1	4	1
2014 年 9 月 19 日	8910100309	2	0	4	0
2014 年 9 月 9 日	8810101463	3	0	1	0
2014 年 9 月 9 日	8710100857	7	0	0	0

***数据详见：/示例程序/data/model.xls**

6.2.5 构建模型

1. 构建窃漏电用户识别模型

在专家样本准备完成后，需要划分测试样本和训练样本，随机选取 20%作为测试样本，剩下的作为训练样本。窃漏电用户识别可通过构建分类预测模型来实现，比较常用的分类预

测模型有 LM 神经网络和 CART 决策树，各个模型都有各自的优点，故采用这两种方法构建窃漏电用户识别，并从中选择最优的分类模型。构建 LM 神经网络和 CART 决策树模型时输入项包括电量趋势下降指标、线损类指标和告警类指标，输出项为窃漏电标识。

1) 数据划分

对专家样本随机选取 20%作为测试样本，剩下的 80%作为训练样本。其代码如代码清单 6-2 所示。

代码清单 6-2 原始数据分为训练数据测试数据

```
%% 把数据分为两部分：训练数据、测试数据
clear;
% 参数初始化
datafile = '../data/model.xls'; % 数据文件
trainfile = '../tmp/train_model.xls' ; % 训练数据文件
testfile = '../tmp/test_model.xls' ; % 测试数据文件
proportion =0.8 ;  % 设置训练数据比例

%% 数据分割
[num,txt]= xlsread(datafile);
% split2train_test 为自定义函数，把 num 变量数据（按行分布）分为两部分
% 其中训练数据集占比 proportion
[train,test] = split2train_test(num,proportion);

%% 数据存储
xlswrite(trainfile,[txt;num2cell(train)]); % 写入训练数据
xlswrite(testfile,[txt;num2cell(test)]); % 写入测试数据
disp('数据分割完成！');
```

2) LM 神经网络

设定 LM 神经网络的输入节点数为 3，输出节点数为 2，隐层节点数为 10，显示间隔次数为 25，最大循环次数为 1000，目标误差为 0.0，初始 mu 为 0.001，mu 增长比率为 10，mu 减少比率为 0.1，mu 最大值为10^{10}，最大校验失败次数为 6，最小误差梯度 1e-7。训练样本建模的混淆矩阵见图 6-8，分类准确率为 94.0%，正常用户被误判为窃漏电用户占正常用户的 3.4%，窃漏电用户被误判为正常用户占正常窃漏电用户的 2.6%。构建 LM 神经网络模型的代码如代码清单 6-3 所示。

代码清单 6-3 构建 LM 神经网络模型代码

```
%% LM 神经网络模型构建
clear;
```

```
% 参数初始化
trainfile = '../data/train_model.xls'; % 训练数据
netfile = '../tmp/net.mat'; % 构建的神经网络模型存储路径
trainoutputfile = '../tmp/train_output_data.xls' ; % 训练数据模型输出文件

%% 读取数据并转化
[data,txt] = xlsread(trainfile);
input=data(:,1:end-1);
targetoutput=data(:,end);
targetoutput = targetoutput+1; % 所有数据都加 1,方便调用 ind2vec

% 输入数据变换
input=input';
targetoutput=targetoutput';
targetoutput=full(ind2vec(targetoutput));

%% 新建 LM 神经网络，并设置参数
net = patternnet(10,'trainlm');
net.trainParam.epochs=1000;
net.trainParam.show=25;
net.trainParam.showCommandLine=0;
net.trainParam.showWindow=0;
net.trainParam.goal=0;
net.trainParam.time=inf;
net.trainParam.min_grad=1e-6;
net.trainParam.max_fail=5;
net.performFcn='mse';

% 训练神经网络模型
net= train(net,input,targetoutput);

%% 使用训练好的神经网络测试原始数据
output = sim(net,input);

%% 画混淆矩阵图
plotconfusion(targetoutput,output);

%% 数据写入到文件
save(netfile,'net');

output = vec2ind(output);
output = output';
xlswrite(trainoutputfile,[txt,'模型输出';num2cell([data,output-1])]);
disp('LM 神经网络模型构建完成！');
```

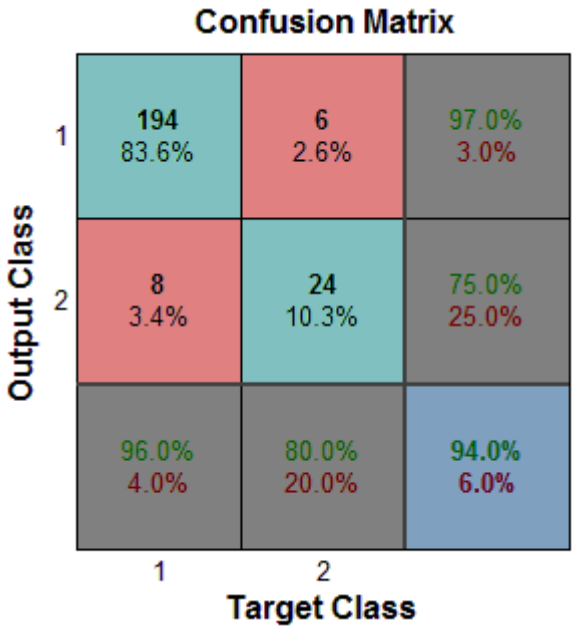

图 6-8 利用训练样本构建 LM 神经网络的混淆矩阵

3) CART 决策树

利用训练样本构建 CART 决策树模型，得到混淆矩阵见图 6-9，分类准确率为 95.3%，正常用户被误判为窃漏电用户占正常用户的 1.3%，窃漏电用户被误判为正常用户占正常窃漏电用户的 3.4%。构建决策树的代码如代码清单 6-4 所示。

代码清单 6-4 构建 CART 决策树模型代码

```
%% 构建 CART 决策树模型

clear;
% 参数初始化
trainfile = '../data/train_model.xls'; % 训练数据
treefile = '../tmp/tree.mat'; % 构建的决策树模型存储路径
trainoutputfile = '../tmp/dt_train_output_data.xls' ; % 训练数据模型输出文件

%% 读取数据，并提取输入输出
[data,txt]=xlsread(trainfile);
input=data(:,1:end-1);
targetoutput=data(:,end);

% 使用训练数据构建决策树
tree= fitctree(input,targetoutput);

%% 使用构建好的决策树模型对原始数据进行测试
output=predict(tree,input);

% 变换数据并画混淆矩阵图
```

```
output=output';
targetoutput=targetoutput';
output= full(ind2vec(output+1));
targetoutput = full(ind2vec(targetoutput+1));
plotconfusion(targetoutput,output);

%% 保存数据
save(treefile,'tree'); % 保存决策树模型

output = vec2ind(output);
output = output';
xlswrite(trainoutputfile,[txt,'模型输出';num2cell([data,output-1])]);
disp('CART 决策树模型构建完成！');
```

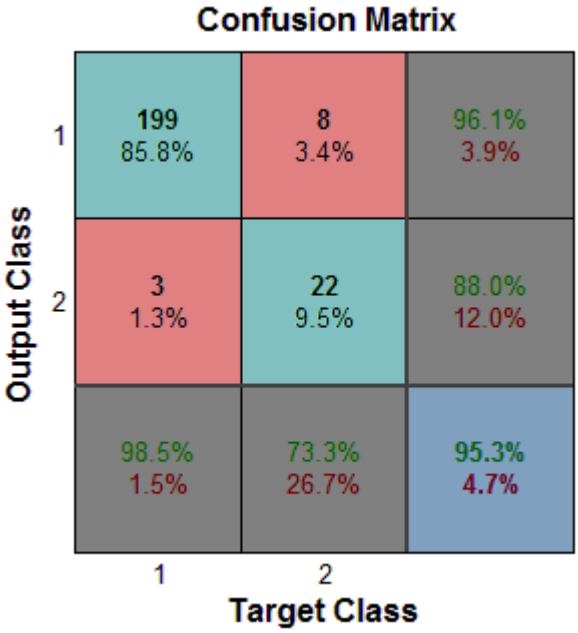

图 6-9 利用训练样本构建 CART 决策树的混淆矩阵

2. 模型评价

对于训练样本，LM 神经网络和 CART 决策树的分类准确率相差不大，均达到 94%。为了进一步评估模型分类的性能，故利用测试样本对两个模型进行评价，评价方法采用 ROC 曲线进行评估，一个优秀分类器所对应的 ROC 曲线应该是尽量靠近单位方形的左上角。分别画出 LM 神经网络和 CART 决策树在测试样本下的 ROC 曲线，见图 6-10 和图 6-11。LM 神经网络和 CART 决策树对测试数据集的测试代码如代码清单 6-5 所示。

代码清单 6-5 LM 神经网络和 CART 决策树测试代码

```
%% LM 神经网络和 CART 决策树模型测试
```

```
clear;
% 参数初始化
testfile = '../data/test_model.xls'; % 训练数据
treefile = '../tmp/tree.mat'; % 决策树模型存储路径
netfile = '../tmp/net.mat';     % 神经网络模型存储路径
dttestoutputfile = '../tmp/dt_test_output_data.xls' ; % 测试数据模型输出文件
lmtestoutputfile = '../tmp/lm_test_output_data.xls' ; % 测试数据模型输出文件

[data,txt] = xlsread(testfile);
input = data(:,1:end-1);
target = data(:,end);

%% 使用构建好的决策树模型对原始数据进行测试
load(treefile); % 载入决策树模型
output_tree=predict(tree,input);

% 决策树输出数据变换以及画 ROC 曲线图
output_tree= full(ind2vec(output_tree'+1));
targetoutput = full(ind2vec(target'+1));
figure(1)
plotroc(targetoutput,output_tree);

%% 使用构建好的神经网络模型对原始数据进行测试
load(netfile);   % 载入神经网络模型
output_lm = sim(net,input');

% 测试数据数据变换以及画 ROC 曲线图
figure(2)
plotroc(targetoutput,output_lm);

%% 写入数据
output_lm=vec2ind(output_lm);
output_lm = output_lm'-1;
output_tree=vec2ind(output_tree);
output_tree=output_tree'-1;
xlswrite(lmtestoutputfile,[txt,'模型输出';num2cell([data,output_lm])]);
xlswrite(dttestoutputfile,[txt,'模型输出';num2cell([data,output_tree])]);
disp('CART 决策树模型和 LM 神经网络模型测试完成！ ');
```

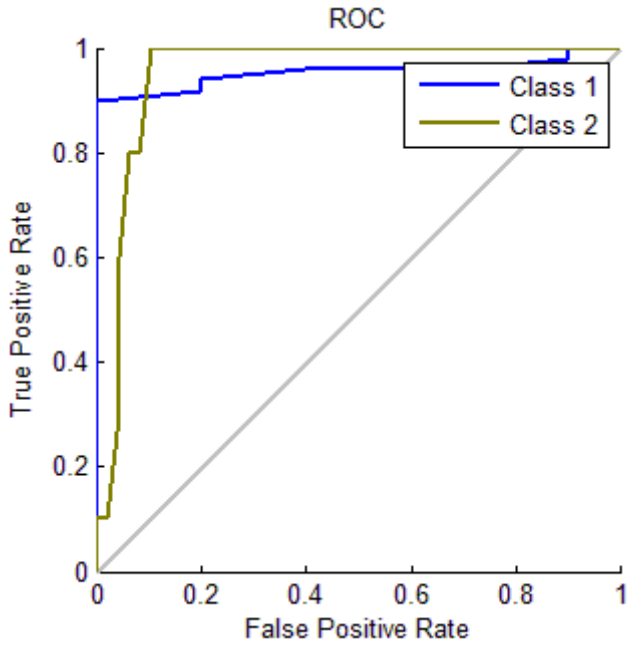

图 6-10 LM 神经网络在测试样本下的 ROC 曲线

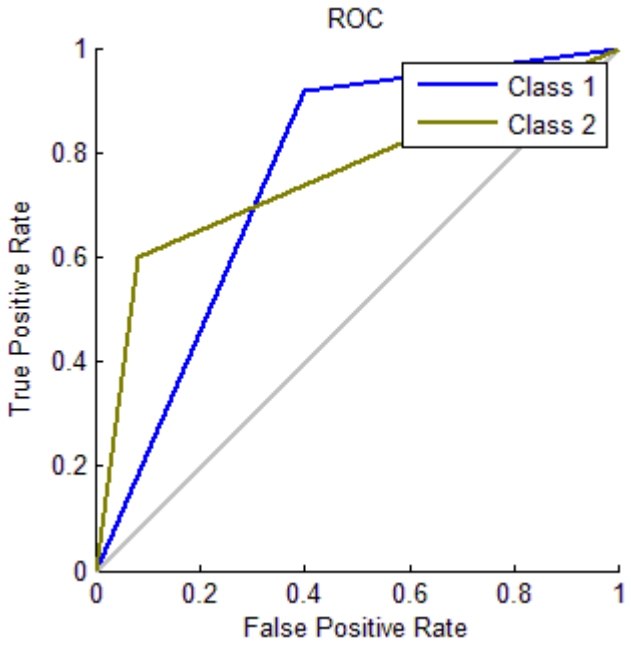

图 6-11 CART 决策树在测试样本下的 ROC 曲线

本案例中主要研究窃漏电用户的识别，所以只观测 LM 神经网络和 CART 决策树 ROC 曲线的 Class2 折线，经过对比发现 LM 神经网络的 ROC 曲线比 CART 决策树的 ROC 曲线更加靠近单位方形的左上角，LM 神经网络 ROC 曲线下的面积更大，说明 LM 神经网络模型的分类性能较好，能应用于窃漏电用户识别。

3. 进行窃漏电诊断

在线监测用户用电负荷及终端报警数据，并经过 6.2.3 节的处理，得到模型输入数据，利用构建好的窃漏电用户识别模型计算用户的窃漏电诊断结果，实现窃漏电用户实时诊断，并与实际稽查结果作对比，见表 6-9，可以发现正确识别出窃漏电用户有 10 个，错误地判断用户为窃漏电用户有 1 个，诊断结果没发现窃漏电用户有 4 个，整体来看窃漏电诊断的准确率是比较高，下一步的工作是针对漏判的用户，研究其在窃漏电期间的用电行为，优化模型的特征，提高识别的准确率。

表 6-9 窃漏电诊断结果与实际稽查结果作对比

客户编号	客户名称	窃电开始日期	结果
7110100608	某塑胶制品厂	2014.6.2	正确诊断
9900508537	某经济合作社	2014.8.20	正确诊断
9900531988	某模具有限公司	2014.8.21	正确诊断
8210101409	某科技有限公司	2014.8.10	正确诊断
8910100571	某股份经济合作社	2014.2.23	漏判
8210100795	某表壳加工厂	2014.6.1	正确诊断
9900287332	某电子有限公司	2014.5.15	漏判
6710100757	某镇某经济联合社	2014.2.21	漏判
9900378363	某装饰材料有限公司	2014.7.6	误判
9900145275	某实业投资有限公司	2014.11.3	正确诊断
8410101508	某玩具厂有限公司	2014.9.1	正确诊断
9900150075	某镇某经济联合社	2014.4.14	漏判
8010106555	某电子有限公司	2014.5.19	正确诊断
7410101282	某投资有限公司	2014.2.8	正确诊断
8410101060	某电子有限公司	2014.5.4	正确诊断

6.3 上机实验

1. 实验目的

- 掌握拉格朗日插值法进行缺失值处理。
- 掌握 LM 神经网络和 CART 决策树构建分类模型。

2. 实验内容

- 用户的用电数据存在缺失值，数据见“/data/missing_data.xls”，利用拉格朗日插值算法补全数据。
- 对所有窃漏电用户及正常用户的电量、告警及线损数据和该用户在当天是否窃漏电的标识，按窃漏电评价指标进行处理并选取其中 291 个样本数据，得到专家样本，数据见“/data/model.xls”，分别使用 LM 神经网络和 CART 决策树实现分类预测模型，利用混淆矩阵和 ROC 曲线对模型进行评价。

注意：数据 80%作为训练样本，剩下的 20%作为测试样本。

3. 实验方法与步骤

实验一

1. 打开 MATLAB 软件，把 missing_data.xls 数据放入当前工作目录。
2. 使用 xlsread 函数把数据读入当前工作目录。
3. 针对读入的数据每一列，进行编程。编程主要参考第 4 章的拉格朗日插值算法，具体主要步骤如下。
   （1） 针对每列数据的每一个缺失值，逐个进行补数（这样可以在连续两个缺失值的情况下，使用前面一个已经补数的值来再次补数后面的一个值）。
   （2） 针对一个缺失值，构造参考组。选取前面 5 个作为前参考组，后面 5 个为后参考组。如果前参考组或后参考组不足 5 个则按实际个数构造参考组。
   （3） 确认缺失值在参考组中的相对位置，然后使用拉格朗日插值进行缺失值插值。
   （4） 根据插值后的值更新原始数据中相应位置的值。
4. 编写并运行程序后，查看插值补数的值是否和给定的参考值一致。

实验二

1. 把经过预处理的专家样本数据 model.xls 数据放入当前工作目录，并使用 xlsread 函数读入当前工作空间。

2. 把工作空间的建模数据随机分为两部分，一部分用于训练，一部分用于测试。
3. 使用 fitctree 函数以及训练数据构建 CART 决策树模型，使用 predict 函数和构建的 CART 决策树模型分别对训练和测试数据进行分类，并与真实值进行对比，得到模型正确率，同时使用 plotconfusion 和 plotroc 函数画混淆矩阵和 ROC 曲线图（这里需要注意 plotconfusion 和 plotroc 函数的输入需要使用 0，1 编码对样本标号进行编码）。
4. 使用 patternnet 函数以及训练数据构建 LM 神经网络模型，使用 sim 函数和构建的神经网络模型分别对训练和测试数据进行分类，参考第 3 步得到模型正确率、混淆矩阵和 ROC 曲线图。
5. 对比分析 CART 决策树模型和 LM 神经网络模型针对专家样本数据处理结果的好坏。

4. 思考与实验总结

1. 在进行插值补数选取参考值时，为什么选择 10 个为一组？
2. 编写 MATLAB 自带的缺失值补数方法和拉格朗日插值补数方法进行对比。

6.4拓展思考

目前企业偷漏税现象泛滥，严重影响国家的经济基础。为了维护国家的权力与利益，应该加大对企业偷漏税行为的防范工作。如何用数据挖掘的思想，智能的识别企业偷漏税行为，有力地打击企业偷漏税的违法行为，维护国家的经济损失和社会秩序。

汽车销售行业，通常是指销售汽车整车的行业。汽车销售行业在税收上存在少开发票金额、少计收入，上牌、按揭 、保险等一条龙服务未入帐反映，不及时确认保修索赔款等多种情况，导致政府损失大量税收。汽车销售企业的部分经营指标能一定程度上评估企业的偷漏税倾向，附件（见：/拓展思考/拓展思考样本数据.xls）提供了汽车销售行业纳税人的各个属性和是否偷漏税标识，请结合汽车销售行业纳税人的各个属性，总结衡量纳税人的经营特征，建立偷漏税行为识别模型，识别偷漏税纳税人。

6.5 小结

本章结合窃漏电用户识别的案例，重点介绍了数据挖掘算法中 LM 神经网络和 CART 决策树算法在实际案例中的应用。研究窃漏电用户的行为特征，总结出窃漏电用户的特征指标，对比 LM 神经网络和 CART 决策树算法在窃漏电用户的识别效果，从中选取最优模型进行窃漏电诊断，并详细的描述了数据挖掘的整个过程，也对其相应的算法提供了 MATLAB 上机实验。