

基于大数据技术的用电信息分析方法研究与应用

基于大数据技术的分时阶梯电价及用电行为分析，可以有效挖掘现有数据中蕴含的价值，不仅能够解决业务发展遇到的瓶颈，而且可以充分发挥信息系统的价值，节约人力、软硬件设备等各项投入。

吕旭明¹ 李钊¹ 雷振江¹ 赵永彬² 杨海峰³ 郑善奇² 杨壮观^{2/1}. 国网辽宁省电力有限公司科技信通部 2. 国网辽宁省电力有限公司信息通信分公司 3. 国网辽宁省电力有限公司后勤工作部



吕旭明/高级工程师

0 引言

辽宁公司根据国网信通部开展大数据应用需求工作的通知，结合辽宁省业务现状，开展了《分时阶梯电价及用电行为分析》的建设工作。依托辽宁公司营销、财务等历史数据，应用基于Hadoop平台的分布式大数据挖掘技术，利用营销业务应用、财务管控系统中海量历史和实时数据，建立基于熵权法的客户信誉综合评价模型，基于决策树模型的电费回收风险评估模型，基于聚类模型的用户用电行为模型。

1 大数据技术研发及应用基础

1.1 技术特征

- (1) 先进性
采用云计算平台，Hadoop，结合数据挖掘算法对数据进行前瞻性分析。
- (2) 安全性原则
数据隔离，采用已有系统的安全性管理组建。
- (3) 可扩展性原则
应采用柔性设计，拥有良好的可扩展

展性，具备业务处理的灵活配置能力，能随着营销辅助决策业务需求变化灵活配置与调整。

1.2 基于Hadoop的分布式存储环境搭建

首先数据采集服务按照预定规则将数据写入数据交换区，数据传输应用程序将交换区数据写入“数据存储区”，该数据区规划在HDFS之上。由“结构化数据抽取应用程序”将结构化指标数据抽取到“结构化分析数据存储区”，并作为HIVE服务的数据存储由HIVE服务将数据以结构化数据服务的形式发布。各模块关系如图1所示。

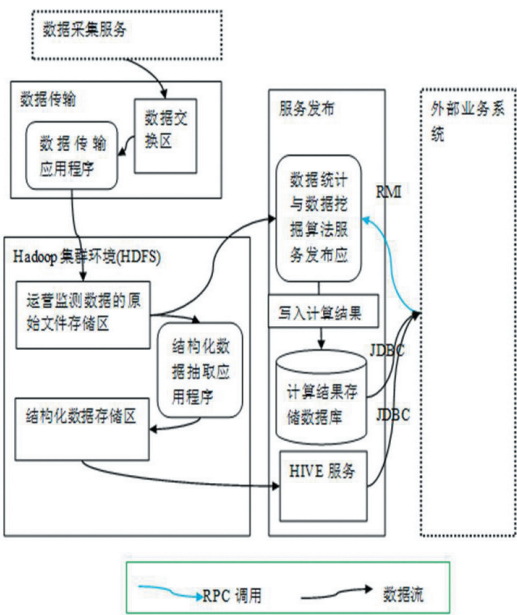


图1 各模块关系图

(1) 容量规划

根据Hadoop平台的存储特点，以3个副本设置Hadoop平台，结合Hadoop集群的可扩展特性，每个节点可规划4 T左右的磁盘即可。

(2) 软件环境设计
软件环境要求如表1所示。

表1 软件环境要求

类别	要求
操作系统	Linux操作系统
数据库	Oracle、mysql
云计算环境	Hadoop、Hive、Mahout

2 基于构建模型的信息分析实现方法

2.1 基于熵权法的用户评价模型

熵权法是一种客观赋权的方法，它的思想是通过计算指标的信息熵，根据指标的相对变化程度对系统整体的影响来决定指标的权重，相对变化程度大的指标具有较大的权重，此方法现广泛应用于在统计学等各个领域，具有较强的研究价值。熵值法首先是计算样本的各指标得分，并利用各指标的信息效用值确定指标的权重，然后求加权 and 得到综合得分。

通过分布式计算得到各项指标的熵值，根据熵值大小确定每项指标的信息效用值。指标的信息效用值越大，对方案评价的重要性就越大，权重就越大。

(1) 模型因素变量选取

电力企业用户评价需要考虑四个方面的因素：用电行为、购电情况、服务成本和信用状况，而每方面又包含若干指标。

用电行为包括：峰时耗电率、谷时耗电率、平段耗电率、各阶梯用电率、负荷率和稳定比率。

购电情况包括：总购电量、最近一次购电量、购电量增长率和平均单价。

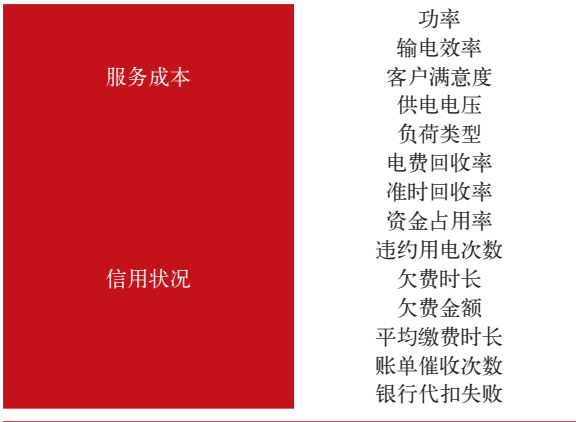
服务成本包括：功率、输电效率、客户满意度、供电电压和负荷类型。

信用状况包括：电费回收率、准时回收率、资金占用率、银行代扣失败、欠费时长、欠费金额、平均缴费时长、违约用电次数和账单催收次数。

四方面评价的层次结构如表2所示。

表2 评价层次结构表

用电行为	稳定比率 负荷率 平段耗电率 峰时耗电率 谷时耗电率 各阶梯用电率
购电情况	总购电量 最近购电量 购电量增长率 平均单价



(2) 样本数据选取

目前辽宁省针对居民用户采取阶梯电价的计价方式，针对非居民用户采取分时电价的计价方式，评价模型中的用电行为包含了用户在不同的价格时段用电情况分布，因此对居民和非居民用不同的评价样本数据，分别建立评价模型。初步拟定选取2011~2014年共4年的数据（年数至少需要2年）。

(3) 建立模型

通过建模流程完成客户评价模型的建模过程。如图2所示。

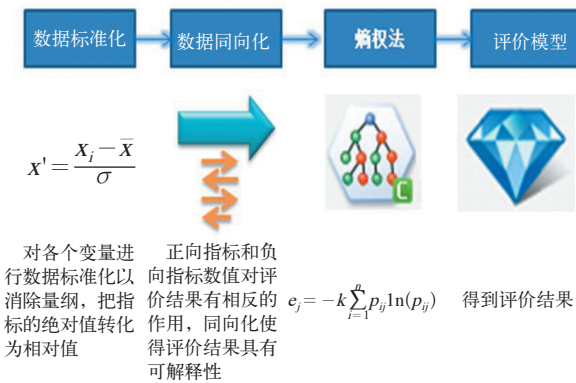


图2 建模流程图

(4) 客户评价模型应用

通过熵权法可以从客观的角度得到各个指标对评价方面的影响权重，从而得到各个用户的整体评价，根据评价结果，针对不同的评价水平，电力企业可以采取不同的电费收缴方式、催收政策，保障企业利益。

2.2 基于决策树模型的客户风险预测

电费回收风险预测，是对用户未来按期缴费行为的提前预测，希望通过差异化提醒和账

单等方式，提高电费回收率，因此需要采用预测模型。

在客户电费回收风险预测中，首先根据营销系统里的与电费回收相关的各字段信息作为多因素输入变量，提取历史数据作为风险样本选取标准，近期数据作为预测数据，运用算法模型进而预测客户的电费回收风险，并在风险值基础上对客户进行风险等级细分。

此处主要难点为不同的供电局，电费回收风险“坏客户”样本选取标准会存在不同的差异，需要结合实际情况进行划分。

(1) 模型因素变量选取

主要提取营销系统里的与电费回收相关的各字段信息，如表3所示。

表3 模型因素相关变量表

字段名称	
客户编码	最近12个月总用电量等级划分
客户名称	最近03个月发生电费实际金额
地市局编码	最近06个月发生电费实际金额
电价代码	最近24个月发生电费实际金额
用电性质	最近03个月缴费方式变更次数
用电类别	最近03个月缴费次数
客户身份	最近06个月缴费方式变更次数
供电电压	最近06个月缴费次数
合同容量	最近12个月缴费方式变更次数
负荷类型	最近12个月缴费次数
当前是否销户	最近24个月缴费方式变更次数
是否存在预购电装置	最近24个月缴费次数
城镇用户/农村用户	最近03个月平均缴费时长
出账周期	最近06个月平均缴费时长
行业大类	最近12个月平均缴费时长
用电行为	最近24个月平均缴费时长
购电情况	最近12个月违约用电次数
服务成本	最近12个月罚金金额
信用状况	最近12个月划扣失败次数
客户类型（居民/非居民）	最近03个月账单催收次数
最近12个月总用电量	最近06个月账单催收次数
最近12个月发生电费实际金额	最近12个月账单催收次数
最近12个月总用电量与去年同期对比	最近24个月账单催收次数

(2) 样本数据选取

并不是所有有过欠费的客户都是我们需要关注的“坏客户”，因为存在一些小金额偶然忘记缴费的客户。对于非居民和居民客户，“坏客户”的标准应该不同，坏客户的比例在0.5%~10%都可以建模，因此坏客户的挑选可以参考更多业务经验。

在此将“坏客户”样本集分为非居民和居

民类别。针对选择标准，居民和非居民均以欠费次数和欠费金额作为评价因素，具体他们的次数和金额应该根据总欠费次数和金额来评定。首先根据总数据的平均值、中位数和75分位数进行选取，然后根据业务专家的实际经验进行选取使用哪种方式能达到更好的效果。

“坏客户”标准制定完成后，在观察期内（样本数据），按坏客户界定标准，分居民、非居民分别计算是否为坏客户。按用户号将预测变量（基本属性及历史行为属性）和目标变量（是否为坏客户）进行拼接，用于最终建模。

(3) 建立模型

通过建模流程完成电费回收风险的建模过程。如图3所示。

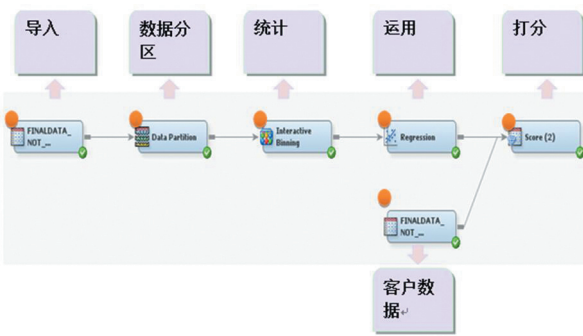


图3 建模流程图

其中，针对建模样本数据，所有非居民中，随机抽取40%样本用于建立模型。所有居民中，随机抽取10%样本用于建立模型。

针对数据分区，样本数据根据建模需要，通过Data Partition节点，按40%，30%，30%拆分成训练集、验证集和测试集。训练集：用于建立回归模型和计算回归系数；验证集：对训练集所建立的模型结果进行验证和细微调整；测试集：用于模型测试。

(4) 客户群体划分

将上一步模型得出的客户风险等级分数从高到低进行排列，可以按照供电局的实际情况，划分不同等级的客户风险度。一般情况下高风险客户群主要集中在前5%。将前5%客户进一步细分，可以确定前1%的纯度非常高。因此建议分群方法为：前1%：高风险群；2~5%：次高风险群；后95%：低风险群。

根据上面划分出的群体，具体分析每个群体的特征，针对特征采取差异化的服务。

2.3 基于聚类模型的用电行为分析

聚类分析又称群分析，它是研究（样品或指标）分类问题的一种统计分析方法。它是由若干模式组成的，通常模式是一个度量的向量，或者是多维空间中的一个点。聚类分析的目的是根据数据的内在特性，将数据划分为若干独立分组，使得每组内部成员之间相似性较大，而和其他组成员之间相似性较小。

在聚类模型中，主要将客户每一年的12个月份的平均分时电价作为输入变量，通过观察每年的电价变化趋势寻找特征出现变化的客户，作为用电异常客户的重点关注对象。

(1) 模型变量选取

此处主要针对用户每个月的分时电价进行聚类，因此主要提取营销系统的用户基本档案信息和每个月的分时电价。

(2) 模型数据选取

在此选取使用分时电价的客户，由于需要分析的是每年中12个月份的用电变化特征，因此需要2011~2014年总共4年的数据（年数至少需要2年），然后将12个月份的电价数据进行纵向化处理。

(3) 建立模型

建模流程图如图4所示。

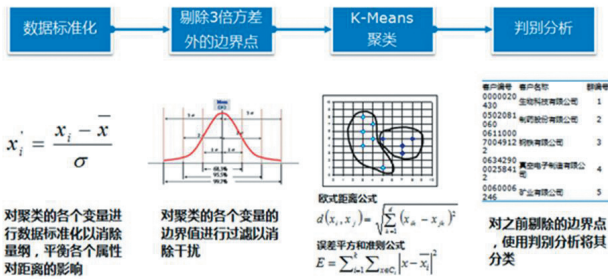


图4 建模流程图

通过以上建模流程，完成分时电价分析的建模过程。

(4) 客户群体划分

根据客户的数据进行聚类之后，每个客户每年会得到一个群体的所属，分析每个群体的数据特征，查找几年之内属于不同群体的客户，将此类客户划分为用电异常客户。

3 结束语

结合营销业务现状，依托辽宁公司营销、财务和生产等历史数据，研究基于Hadoop平台的分布式大数据挖掘技术，探索客户信誉综合评价模型、电费回收风险评估模型和用户用电行为模型等方面的大数据原型系统，有效挖掘现有数据中蕴含的价值，能够解决业务发展遇到的瓶颈，可以充分发挥信息系统的价值，节约人力和软硬件设备等各项投入。

1) 基于大数据技术，利用营销基础信息平台中海量数据，建立客户评价模型，跟踪分析单个客户或行业客户的付款历史和习惯。

2) 开展电费回收风险分析，保障企业的电费收入，有效控制企业经营风险。

3) 对用户用电行为进行聚类，从海量用户中分离出相似用户，识别异常用户和异常用电行为。

参考文献

[1] Splitting an existing replicat with the RANGE command into multiple replicats. (Doc ID 1320133.1)

[2] 中国电机工程学会《中国电力大数据发展白皮书》[S].

[3] 《关于加强中央企业信息化工作指导意见》（国资发〔2007〕8号）[S].

[4] 刘军. 基于云计算的网络数据灾备关键技术的研究[D]. 武汉: 华中师范大学, 2011.

[5] 杨晓红, 李健, 杨卫国. 信息系统容灾技术的分析与研究[J]. 计算机工程与设计, 2005, (10): 2727-2729.

[6] 中国国家标准化管理委员会《GB/T 11457-1995 软件工程术语》[S].

作者简介

吕旭明（1981-），男，硕士研究生，高级工程师，主要研究方向为电力企业信息化，智能电网，信息安全。

（收稿日期：2015-12-10）EA
（编辑：崔文静）