

电力用户行为画像构建技术研究

电力行业数据涵盖了丰富的用户行为信息，但仍存在数值型数据比重大且很难被业务人员直观理解的问题。针对电力行业的特性提出一个面向用户行为的标签层次体系，以及基于密度、距离的聚类算法和梯度提升树算法的标签生成方法，形成了一整套电力用户行为画像技术。通过实际数据验证，该行为画像技术有效地解决标签语义化问题，方便识别用户行为的发生，并通过刻画用户行为的统计特征，预知行为发生的可能性，面向用户行为的用户分类，最终完整刻画一个电力用户。

傅 军¹ 许 鑫¹ 罗 迪² 朱天博¹ 刘 霞³/1.国网冀北电力有限公司电力科学研究院 2.北京科技大学计算机与通信工程学院 3.北京博望华科科技有限公司

0 引言

电力用户行为多种多样，包括购电、缴费、业扩报装、容量变更及违约用电等。电力企业获取了大量的电力用户的业务行为信息，亟待将其转化为价值，这需要对电力用户画像，用以分析用户的消费特征、行为偏好，进而为大数据分析挖掘奠定基础。

用户画像的本质是用户特征的标签化，包括用户画像标签体系和标签生成技术。电力行业对于用户的分析主要是基于行为的，而行为数据大多为数值型数据，需要经过一定的转化规则才能转化为便于业务人员理解的语义标签。且电力行业的各种业务行为是商业行为，遵循客观经济规律，可进行较精细的预测。

在互联网用户画像构建技术方面前人做了很多研究，Ali等人在视频录制领域根据用户的录制记录，使用协同过

滤矩阵构建了用户画像系统，名字叫做TiVo^[1]；Slimani等人在在线教育领域通过创建资源的索引，使用动态贝叶斯算法为用户画像^[2]。目前在电力行业中，关于电力行业用户画像构建技术的文献还较少。林森、欧阳柳提出了一套基于大数据理论的电力客户画像标签体系^[3]，从业务的角度分析了标签体系的维度和内容，以及标签库的管理框架。但他们并没有具体讨论生成画像的技术和方法，在整个行业内也缺乏用户画像生成技术的讨论，缺乏可以借鉴的其他用大数据技术对电力用户进行分析的研究。周李、赵露君等人提出了利用稀疏编码模型发掘电力用户的异常用电行为^[4]，可用于构建电力用户的异常用电画像；涂莹等人针对高压用户、低压非居民用户和低压居民用户构建基于逻辑回归的用户电费回收风险预测模型^[5]，可用于刻画用户欠费风险，但仍存在过拟合、难理解及可解释性差的问题。

本文提出一整套适用于电力行业的用户行为画像技术，包括用户标签层次体系和标签生成方法，将数值型数据转为易于理解的语义行为标签，方便识别用户行为的发生，统计用户行为偏好等特征，预知行为发生的可能性，并根据用户行为特征对用户分类，从用户行为的角度去刻画一个用户。

1 电力用户标签体系

1.1 电力用户标签体系构成

电力用户行为画像主要包含电力



傅 军/高级工程师

关键词/Keywords

行为画像 ·
行为预测 ·
聚类 ·
梯度提升树 ·
电力行业 ·

基于大数据的客户增值服务平台关键技术研究及应用
(52018K15000G)。

用户两类重要信息：静态属性信息和动态行为信息。静态属性信息为电力用户较为稳定的信息，如电压等级、用电规模和行业等。动态行为属性即用户不断变化的行为信息，例如增容行为、违约行为及缴费行为等，这些行为的发生时间和行为变化量是不断改变的。要将电力用户数据转为商业价值，电力企业需要识别电力用户某一行为的发生、描述该行为的时间特征及发现各个行为之间的相关性，进而根据行为对电力用户进行分类和行为预测。因此，提出由用户基本属性标签、行为标签、行为描述标签、行为预测标签和分类标签组成的电力用户画像的用户标签体系，标签之间的关系如图1所示。

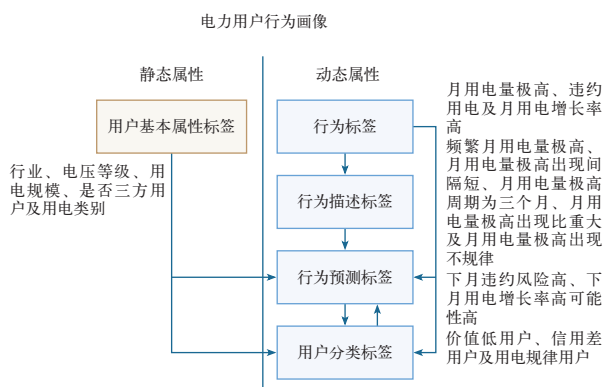


图1 电力用户行为画像标签体系

1.2 行为标签

电力企业的数据多为数值型数据，无法直接体现用户行为语义特征；领域专家可依据一定的规则和已有多种聚类算法将数值属性数据进行离散化，划分成几个易于理解的语义标签层级，把定量数据转化为定性行为语义标签 T ； T 是一个三元组，定义如下

$$T = \langle t, N, n \rangle \quad (1)$$

式中， t 为标签形成时间； N 为标签名； n 为行为状态。行为语义标签 T 表示了用户某个时刻发生了什么业务行为以及行为状态。

1.3 行为描述标签

行为描述标签体现用户行为的时间特征、行为偏好。每个行为标签代表的行为特征可以用频次、平均值、覆盖率、偏离度、平均时间间隔、

周期特征及时段偏好特征来表示。覆盖率表示某个时间段内该业务行为出现的次数占同类业务行为出现次数总和的比重。行为出现平均时间间隔即行为标签出现时间间隔的平均值。偏离度即行为标签出现的时间间隔的标准差，体现用户产生某种行为的时间均匀程度，偏离度越低，则该行为有可能是一种周期性行为。周期性用来衡量用户某行为是否具有周期性。时段偏好特征表示用户行为产生的时间段偏好。上述特征均从历史态和近态来描述，突出某行为的时间特征。

这些特征共同描述了用户行为的统计特征，有效刻画用户的某个行为，并作为使用机器学习模型预测用户行为发生和对用户进行分类的输入。

1.4 行为预测标签与分类标签

行为预测标签用于描述电力行业某业务行为在未来发生的可能性。电力业务行为的随机性较小，与其他业务行为和用户基本属性具有强关联关系，可以实现较准确的预测。对于业务行为的预测可以让电力企业合理安排生产计划以及提供个性化服务，如预测电力用户转化为“费控”用户的可能性。分类标签是从多个角度根据电力用户的行为，使用聚类或者专家经验法将电力用户划分为多个种类，如价值高、中和低的用户，信用高、中和低的用户，从而描述电力用户的价值画像和信用画像。

2 标签生成技术

2.1 基于密度和距离聚类的行为标签生成方法

传统K均值（K-means）聚类方法必须先知道确定的聚类中心数目，并不能很好处理非球状分布的数据^[6]。在生成行为标签时，需要进行单维度的聚类。本文使用Alex Rodriguez和Alessandro Latio在Science上提出的聚类思想对电力行业的数据进行聚类，能有效解决以上问题，其核心是利用局部密度和距离找出数据点的聚类中心。

待聚类的数据集定义为 $S = \{X_i\}_{i=1}^N$ ， $I_s = \{1, 2, 3, \dots, N\}$ 为相应的指标集，表示数据集中两个

点 X_i 和 X_j 的距离。对于 S 中的每一个点 X_i ，计算其局部密度 ρ_i 和距离 δ_i 。首先定义 d_c 为截断距离，需要事先指定。 ρ_i 的计算公式如下

$$\rho_i = \sum_{j \in I_S} \varphi(d_{ij} - d_c) \quad (2)$$

其中

$$\varphi(x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \quad (3)$$

假设 $\{q_i\}_{i=1}^N$ 为 $\{p_i\}_{i=1}^N$ 的一个降序排列下标序，即 $\rho_{q_1} \geq \rho_{q_2} \geq \dots \geq \rho_{q_N}$ ，则 δ_{q_i} 计算公式如下

$$\delta_{q_i} = \begin{cases} \min(d_{q_i q_j}) & i \geq 2, j < i \\ \max(\delta_{q_j}) & i = 1, j \geq 2 \end{cases} \quad (4)$$

假设一个待聚类的数据集 S 有 n_c ($n_c > 0$) 个聚类中心。 $\{m_i\}_{i=1}^N$ 为各个聚类中心的编号； $\{c_i\}_{i=1}^N$ 为数据点归类属性标记； $d_{\max} = \max\{d_{ij}\}$ 为 S 中距离最远的两个点的距离； $\{n_i\}_{i=1}^N$ 中的 n_i 为 S 中所有局部密度比 X_i 大的数据点中与 X_i 距离最近的数据点的编号。

聚类算法步骤如下所述。

(1) 初始化和预处理

1) 给定用于确定截断距离 d_c 的参数 t ， $t \in (0, 1)$ 。

2) 计算数据集 S 中各个数据点相互之间的距离 d_{ij} 且 $i > j$ ， $i, j \in I_S$ 。

3) 由2) 可得 $M = 0.5(N-1)N$ 个距离，对其进行升序排列得 $d_1 \leq d_2 \leq \dots \leq d_M$ ，令 $d_c = d_{f(M)}$ ， $f(M)$ 为对 M 进行四舍五入后的整数。

4) 按照式 (2) 计算每个数据点的局部密度 $\{p_i\}_{i=1}^N$ ，并进行降序排序得到 $\{q_i\}_{i=1}^N$ 下标。

5) 按照式 (4) 计算每个数据点的距离，并生成 $\{n_i\}_{i=1}^N$ 。

(2) 确定聚类中心 $\{m_i\}_{i=1}^N$ ，并初始化 $\{c_i\}_{i=1}^N$

聚类中心的确定必须符合两个条件：① 聚类中心本身局部密度大，即被密度不超过它的邻居包围；② 与其他密度更大的数据点间的距离更大。即对于每个数据点 X_i ，其 ρ_i 和 δ_i 越大，越有可能成为聚类中心。计算 ρ_i 和 δ_i 综合考虑的量，令

$$\gamma_i = \rho_i \delta_i \quad i \in I_S \quad (5)$$

显然， γ_i 值越大，越有可能是聚类中心。画出 γ_i 降序排列的示意图，必定存在人工能够识别出来

的 γ_i 值较大的点，将其作为聚类中心。

(3) 对非聚类中心数据点进行归类

对非聚类中心归类时，是按照 ρ_i 的值从大到小的顺序进行的，根据 $\{n_i\}_{i=1}^N$ 可知局部密度比该数据点大的且距离最近的数据点的编号，由算法思想可知，该数据点必定和与其距离最近的且局部密度比其大的数据点属于同一个簇。

2.2 行为描述标签生成方法

用户 u 在某个时间段用户行为标签 at 的覆盖率用CreateRatio表示， $\text{sum}(at_j, u)_{ET-ST}$ 为用户 u 某类业务行为的行为标签 j 在某个时间段发生的次数总和， $ET-ST$ 为统计的时间长度，则覆盖率计算公式可表示为

$$\text{CreateRatio} = \frac{\text{sum}(at_j, u)_{ET-ST}}{\sum_{i=0}^n \text{sum}(at_i, u)_{ET-ST}} \quad (6)$$

行为标签出现的时间间隔用 d 表示，则行为出现平均时间间隔计算公式为

$$\text{average}(d) = \frac{\sum_{j=0}^n d_j}{k} \quad (7)$$

偏离度计算公式如下

$$\text{BalanceRatio}_{at_j, u} = \sqrt{\frac{\sum_{i=0}^n [(d_i - \text{average}(d))^2]}{k}} \quad (8)$$

如果某行为没有周期性，则用0表示，若有则用其周期时间间隔 d 表示。设一个用户某行为标签在某段时间内出现的时间距离为 $d_1, d_2, \dots, d_k$ ，对时间距离划分出 n 个区间，形成 n 个等级 $d_1, d_2, \dots, d_n$ 。如果 $\text{sum}(d_i, u)$ 表示用户 u 的行为间隔 d_i 出现的次数，则周期性 $\text{period}(at, u)$ 的计算公式为

$$\text{period}(at, u) = \begin{cases} d_i & \begin{cases} \exists d_i \in (d_1, d_2, \dots, d_n), \text{sum}(d_i, u) \geq 0.6, \\ \sum_{j=0}^n \text{sum}(d_j, u) \end{cases} \\ \forall d_i \in (d_1, d_2, \dots, d_n), \text{sum}(d_i, u) < 0.6, \\ 0 & \sum_{j=0}^n \text{sum}(d_j, u) \end{cases} \quad (9)$$

如果用户某行为发生在该时间段的次数占行为发生总次数的60%，则该行为具有时间段偏好，时间段偏好特征 TF 即为该时间段。

对行为的各个时间特征进行单维度的离散化，对行为的各个时间特征划分层级；对行为的频次、覆盖率及平均时间间隔进行聚类，根据专家经验定义每个类簇，描述行为的偏好；对行为的偏离度聚类得到的标签用于描述行为的时间间隔均匀程度；使用时段区间名称作为行为时段频率特征的标签。

2.3 基于梯度提升树的行为预测标签生成方法

电力用户画像的构建涉及到用户行为预测的问题。本文主要采用梯度提升树算法来构建行为预测模型。Gradient Boosting其实是一个框架，可以套入各种不同的分类算法，通过不断迭代改进上一次分类结果来提高分类准确度^[7]。传统的Boosting方法是在算法开始的时候，为每一个样本初始化一个相同的权重，对训练样本进行迭代，每一次迭代结果提高分类错误训练样本的权重，降低分类正确的训练样本权重，进行 N 轮迭代后得到 N 个分类器，最后将其组合起来。Gradient Boosting的每一次计算是为了减少上一次的残差，通过在残差减少的梯度方向上建立一个新的分类器。以下为GBDT算法的主要步骤^[8]。

1) 输入： $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, $x_i, y_i \in R$ 。 T 为训练样本数据集， x 为自变量， y 为因变量， N 为数据集大小。

2) 初始化

$$f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma) \quad (10)$$

3) 对 $m = 1, 2, 3, \dots, M$, M 为迭代次数。

a) 对 $i = 1, 2, 3, \dots, N$, 计算

$$r_{im} = - \left[\frac{\partial L[y_i, f(x_i)]}{\partial f(x_i)} \right]_{f=f_{m-1}} \quad (11)$$

式中， r_{im} 为梯度方向。

b) 对 r_{im} 拟合一个回归树，得到树 m 的叶节点区域 R_{jm} , $j = 1, 2, \dots, J_m$ 。

c) 对 $j = 1, 2, 3, \dots, J_m$, 计算

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L[y_i, f_{m-1}(x_i) + \gamma] \quad (12)$$

d) 更新

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \quad (13)$$

4) 得到回归树

$$f(x) = f_M(x) \quad (14)$$

用户分类标签是在以上三类标签生成的基础上通过聚类的方法得到的，将用户分成不同的有意义的几类。

3 实验与分析

本文采用国网冀北电力有限公司真实的大工业用户数据作为样本，共20 911户。该数据集还包含从2014年1月~2016年6月的月用电量数据183 662条，13 839条容量变更数据，23 562条用电检查数据，20 911名大工业用户2014年1月~2016年6月的违约记录和基本信息。

3.1 行为标签生成

对用电行为数据进行单维度离散化生成行为标签，给电力用户每月的用电行为打上用电量极高、用电量高、用电量中、用电量低和用电量极低五个行为标签。按照上文的聚类方法，根据多次试验，设置 $t = 0.05$ 时聚类效果最好。计算每个数据点的局部密度和距离后，又分别计算每个数据点的 γ 值，进行降序排列如图2所示。

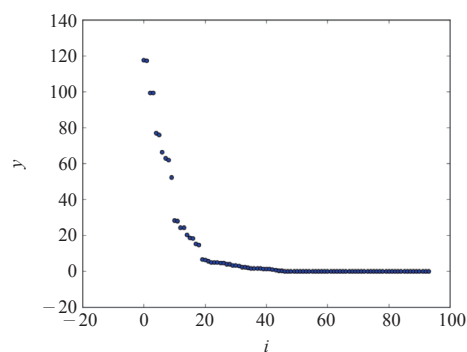


图2 用电量的降序排列的 γ 值示意图

由图2可见，选择 γ 值前20的点作为聚类中心，根据聚类中心大小分别分成用电量高、用电量中、用电量低和用电量极低，然后将非聚类中

心的数据点进行归类。因此，最终可将电力用户的每月用电量分为五个等级，并建立相应的行为标签。电力用户月用电量数据还可以按照月增长量，划分出用电增长幅度大、用电变化幅度小和用电略微减少等行为标签；其他行为的划分使用

相同方法可得。

3.2 行为描述标签生成

本文以电力用户每月用电行为和容量变更行为的行为描述标签生成为例。根据上文的公式计算某用户对应行为的行为描述标签，结果如表1所示。

表1 部分行为特征

行为	频次	覆盖率	平均间隔	偏离度	周期	时段偏好
没有用电	17	0.57	1.5个月	1.36	1个月	上半年
用电量低	7	0.23	2.1个月	2.06	无周期性	下半年
用电量极低	4	0.13	2.75个月	3.05	无周期性	无明显偏好
暂停	5	0.56	6个月	3.36	无周期性	年初
暂停恢复	4	0.44	7个月	0	7个月	下半年

由表1，根据上文的聚类算法将其离散化，可得到该用户的部分行为描述标签：频繁停止用电、月用电量低、容量暂停集中在年初和暂停恢复周期为7个月。

3.3 行为预测标签、分类标签生成

如表2所示，采用上文提到的聚类算法和GBDT算法分别构建违约用电风险分类标签。利用从RFM模型改进的RFMO模型来定义用户的违约用电风险度。

表2 传统RFM模型和RFMO模型

模型	R	F	M	O
传统RFM	客户近一次购买距分析点的时长	客户一定时期内购买的次数	客户一定时期内购买的总金额	无
RFMO	客户近一次违约用电距分析点时长	客户一定时期内违约用电次数	客户一定时期内违约用电总电量	客户一定时期内违约用电电量占比

本文分别计算电力用户近9个月违约总次数、近9个月中近一次违约至今的月份数、近9个月违约总电量、近9个月违约总电量与9个月总用电量的占比。对大工业用户的R、F、M和O值进行聚类，最终可以将用户的违约风险度划分为17个类簇。对于R、F、M和O分别赋予相同的权重，最终将用户的违约风险划分为四类，形成关于违约风险的用户分类标签，分别给用户打上违约风险高、违约风险中、违约风险低和无违约风险的

标签。

根据用户的基本属性、月用电量记录、违约记录和容量变更记录构建基于GBDT的违约用电行为预测模型，不同违约风险分别用0、1、2和3表示。训练指标：近9个月总用电量，上月份的电量，客户立户时长，近6个月总用电量，近3个月总用电量，近9个月电费，近6个月电费，近3个月电费，……，合同容量，用电类别，运行容量。

预测结果采用如表3所示的混淆矩阵。

表3 预测结果混淆矩阵

结果	无	低	中	高
无	363	0	0	0
低	0	589	126	3
中	0	196	588	23
高	0	3	32	70

准确率为0.807 827 396。将用户特征指标数据输入模型，根据预测模型输出的不同分别打上不同的违约风险标签。

3.4 电力用户行为画像应用

电力用户行为画像技术主要应用如下：①用户的快速筛选。电力用户画像是对原本数值型数据的语义化，是标签型数据，大大简化了筛选条件。②监控用户行为，并提出建议。如对进行高压增容但随即进行减容的用户行为建立预测模型，对该类用户提出相应的建议，降低

用户运营成本。③精准化营销。构建电力用户从普通用户转为费控用户的行为预测标签,做到有针对性地向用户提供奖励措施使其转为费控用户。④合理安排,提高服务质量。如统计“偏好营业厅缴费”的用户数量,对这部分用户中具有“用户价值高”标签的用户优先安排缴费,减少排队缴费时间。⑤提供电力信用报告。电力用户画像描述了电力用户的用电行为、违约行为等与信用强相关的行为信息,可直接转化为信用报告。

4 结束语

本文提出的电力用户画像技术能有效识别出电力用户在使用电力企业提供的服务过程中产生的行为,并有效描述和预测各个行为,最终全面、准确地描述一个电力用户。本文也存在不足之处,在标签生成的过程中仍然需要依靠专家的主观经验,对于行为的预测和用户的分类需要不断地对模型进行修正。此外,本文采用的用电行为数据以月为单位,跨度较大,在未来将结合日用电负荷数据、天气及财务等外部数据构建更精细的电力用户画像,以便针对需求提供更多的个性化服务。

参考文献

- [1] Ali K, Van Stam W. TiVo: making show recommendations using a distributed collaborative

filtering architecture [C]. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004: 394-401.

- [2] Slimani H, Faddouli N E, Bennani S, et al. Models of digital educational resources indexing and dynamic user profile evolution [J]. International Journal of Emerging Technologies in Learning, 2016, 11 (1): 26.
- [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] 周李, 赵露君, 高卫国, 等. 稀疏编码模型在电力用户异常用电行为探测中的应用研究 [J]. 电网技术, 2015 (11): 3182-3188.
- [5] 涂莹, 林士勇, 欧阳柳, 等. 基于市场细分的逻辑回归模型在电费回收风险预测中的应用研究 [J]. 电力需求侧管理, 2016, 18 (4): 46-49.
- [6] Xu R, Nd W D. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks, 2005, 16 (3): 645-678.
- [7] 邓仙荣. 基于梯度提升回归算法的O2O推荐模型研究 [D]. 淮南: 安徽理工大学, 2016.
- [8] Friedman J H. Greedy function approximation: a gradient boosting machine [J]. Annals of Statistics, 2000, 29 (5): 1189-1232.

(收稿日期: 2017-03-29) **EA**

欢迎订阅 欢迎投稿 欢迎刊登广告