

电力营销数据聚类方法研究

在大数据被广泛赋予变革意义的时代，电力营销数据也成为电力企业智能化过程中的重要财富。针对电力营销数据进行了聚类方法研究。首先进行了电力数据特征分析，以及对电力数据进行数据挖掘时需要关注的应用要点进行说明。在此基础上，探究了基于网格的电力数据聚类分析，在异常点阈值和聚类阈值分析设定的前提下，通过网格聚类方法，聚类得出用户购买电量和购买频次的主要分布，为电力营销决策提供支持。

慕福林 王刚 王芳 陈俐冰/国网重庆市电力公司

0 引言



慕福林/高级工程师

在数据信息飞速发展的今天，各行各业都开始进行大数据分析研究，以提供高效信息处理以及基于数据的智能决策分析。电力营销作为连接电力上游以及用户的核心组成部分，积累了大量的电力销售以及用户用电行为数据。对这些数据进行专业的数据挖掘分析，建立合理的分析模型越来越成为电力智能化的核心任务之一。

通常针对电力营销的数据分析更多的需要考虑营销数据的多样性和复合性。进行数据分析的过程中，如何有效地识别孤立点，如何有效地进行脏数据清理，以及考虑电力系统数据的时序性问题等，都是需要在营销数据挖掘中考虑的实际问题。在此基础上，通过对营销数据进行有效的聚类和分类分析，可以更好地识别用户群体、获取分类用户用电特征，为后续的电力营销智能化提供基础。

关键词/Keywords

电力营销·
聚类·
网格·
数据挖掘·

1 研究背景

1.1 电力营销数据特征

电力营销系统数据来源广泛，通过不同电力设备进行的数据采集、电力销售数据以及地理信息系统等都在不断地为电力营销系统提供数据。电力营销系统本身已经存储了大量的数据，通过对这些数据进行收集、处理来支撑我国电力系统的高效运行。电力营销数据具有如下特征。

(1) 数据种类多

营销系统所包含的数据既有用户信息，也有销售数据，同时还要针对电力专业的电压、电流和功率等多种指标性数据。不同数据具有不同的表示类型，同时多种数据的组合逻辑也对数据存储和挖掘分析提出挑战。

(2) 数据量大

电力营销系统涉及到多个站点，万级用户数据，同时各个部分还在不断地进行数据采集和动态输入，这种情况下，数据量在不断的增长。

(3) 时序性问题

电力营销数据具有较强的时序性，对电力需求进行预测，对用电特征进行分析，都需要利用数据中的时序标签进行聚类、分类以及相关的关联分析。

1.2 电力营销中的数据挖掘分析

(1) 数据分类在电力营销中的应用

该方法是根据设定的特征维度，对大量数据进行基于多维度特征值的分析，在设定的阈值内将数据进行分类。例如：针对用户用电特征，包括用电时段、每月电价和用电功率等数据，可以将用户进行分类，在此基础上可以进行电价指导等方案的制定。

(2) 聚类算法的意义

相对于分类算法，聚类算法是一种非监督学习方法，没有设计明显的物体或信息分类特性。该方法作为分类问题的统计型方法，在电力数据中使用可以更好地适应大数据特性，具有较好的伸缩性，同时可以适应不同的属性。常见的聚类算法有基于模型的聚类算法、K-means聚类方法和基于网格的聚类方法等。

(3) 时序数据挖掘

时序挖掘可以利用过去的数据结合时间标签进行分时段数据特征分析，同时可以进行未来一段时间的数据趋势预测。针对电力营销系统，可以对电力经济特性进行预测，对电力需求进行预警和处理。

(4) 关联分析方法

关联分析可以将电力交易数据、客户数据和其他采集数据作为数据基础进行项目集合或者对象之间的相关性分析。分析结果还可以包括数据之间的频繁模式、关联规则和因果关系等。可以对电力客户特征和用电量、用电时段之间的关系进行量化分析，通过其中一个因素可以自动分析另外的因素，提供决策参考。

2 典型聚类算法分析比较

聚类分析就是按照特征之间的相似性将数据归结到不同的类别中，即“物以类聚”。聚类分析的描述：假设一个空间维度为 m ，将其中的 n 个向量归为 S 个类中的一个，使得每个向量与聚类中心的距离最近。 m 为样本属性数； n 为样本数量； S 是设定值，也可以是自动分类形成的类数。

2.1 主要数据类型

聚类算法具有两种典型的数据结构。

(1) 数据矩阵

用 p 个变量对 n 个对象进行表示，比如对于用户电力数据，可以用地址ID、月用电量、峰值电量和用电时段等属性来表示，这样的关系可以用 $n \times p$ 的矩阵来进行表示，把这种矩阵叫做数据矩阵。

(2) 相异度矩阵

使用距离 $d(i, j)$ 来对两个对象的差异度进行表示，默认为非负值。该值越小，说明两个对象

越接近；越大，说明两个对象差异越大。

2.2 相似度计算

(1) 序数型变量

对于序数型变量，值的实际大小不重要，变量值域要进行映射，映射范围为 $(0, 1)$ 。

(2) 比例型变量

例如指数型变量，指数即为其度量值。这类变量，通常需要进行去对数。

(3) 连续型变量

对于此类变量，用什么样的单位或者多大区间进行数据分割对分类效果影响较大。通常需要对数据进行标准化处理，用距离单位来表示数据。欧氏距离是常用的方法。两个向量之间的欧氏距离为

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2} \quad (1)$$

2.3 聚类方法分析

(1) 基于模型的聚类方法

基于模型的聚类方法包括两类，一种是统计学方法，一种是神经网络方法。该方法将一个聚类假设成一个具体的模型，然后基于模型进行聚合。在聚类过程中，同时可以自动进行聚类数目设定，通过噪声清晰和孤立点过滤来提升聚类效果。

(2) 基于密度的聚类方法

基于密度的聚类主要针对的是通过对象距离聚类时，智能发现球状簇的现象。该方法针对一个已知数据，对某一个区域数据点个数有具体要求。

(3) 基于网格的聚类方法

网格方法即把空间按照向量维度来进行等长分割，这样空间变为分别具有维度特征的单元格。认为在一个单元格中的为同类别对象，在聚类计算过程中，同一网格的点做整体计算，以提高算法效率。

3 基于网格的聚类算法分析

该部分在前文基本概念的基础上，分析基于网格的聚类算法，以及该算法在电力用户分类时的具体使用效果。

3.1 基本定义

(1) 定义1：类别

设数据集 S 中的每个元素为 S_i ($i = 1, 2, \dots$)

3, …, n), 对于任何一个 S 中的元素, 总存在一个与之距离小于 L 的元素, 那这个数据集 S 就称为类。

(2) 定义2: 相邻网格

两个 m 维网格 W_1 、 W_2 , 若还有另外一个网格分别与这两个网格相邻, 则 W_1 、 W_2 为相邻网格。

3.2 算法实现

(1) 数据结构

1) 区域表。该表中放的是聚类后的空间区域描述数据。

2) 类数量表。表示每个聚类结果中具有的对象数量。

3) 方格表。记录每个方格的具体属性。

(2) 输入

聚类对象数据表, 聚类阈值 L , 异常点阈值 D 。

(3) 输出

分类表, 包括分类个数以及各类中的数据量。

(4) 方法

1) 对 m 维进行等距离分割, 将空间分为多维网格。以每个网格的中心向量为点, 进行网格编号, 并对网格内的记录数量进行统计。

2) 对于对象距离超出异常点阈值的网格, 标记为异常点。

3) 对于非异常网格, 分别计算两两之间的欧氏距离, 计入二维数组 $Dis(i, j)$ 中。

4) 遍历二维数组 $Dis(i, j)$, 找出距离最大的两个网格, 作为两个初始分类。

5) 通过聚类分类方法, 将其他非异常网格进行分类。

6) 最后进行网格合并。

3.3 算法分析

该算法具有以下优点。

1) 该算法可以有效地用于分散多维空间的分类问题。

2) 初始点选择以最远距离的数据为标准, 防止局部最优问题出现。

3) 相比于K-means聚类算法, 该方法的聚类点数是自动生成的, 避免人为设定时造成的假设误差。

4) 异常点发现容易, 且易于剔除。

5) 最终的区域表以聚类网格进行合并得出,

符合现实数据的聚集特征。

4 基于网格的聚类算法试验

4.1 数据集

由于本文主要研究聚类算法在电力营销中的应用, 所以采用的数据集是电力营销相关数据, 主要包括客户特征数据和购电数据。在对10 000多条数据进行清理和随机抽取后, 使用其中的8 000条作为本算法试验数据集。部分客户数据信息如表1所示。

表1 部分用户数据

编号	收入/万元	孩子个数/个	教育	购买电量/kW4-h	购买次数/次
2	5~7	1	硕士	8.05	12
5	7~9	0	硕士	7.53	7
19	5~7	2	硕士	6.24	40
38	1~3	3	高中	6.75	15
47	3~5	0	硕士	6.15	17
55	5~7	0	硕士	6.43	7
96	3~5	2	本科	5.30	5

由表1可以看出, 本数据集每个对象有5个维度, 除次数和电量外, 其他数据属于枚举型数据。在电力营销数据挖掘的过程中, 将购买量和购买次数作为主要分析元素来对用户进行聚类。具体分类个数由算法自动生成。

4.2 试验结果分析

通过分组合试验方式, 以最终异常点数量为取舍标准, 对异常点阈值和聚类阈值进行尝试和选择。阈值选择参考如表2所示。

表2 阈值选择参考

异常点阈值	聚类阈值	累个数/个	异常点个数/个
19	5	15	1 158
20	10	5	2 369
15	5	11	1 475
15	10	5	1 732

从表2中可以看出, 异常点数量与阈值设定值成正相关, 此时, 需要结合样本数据集, 考虑有效试验数据数量, 选取合适的组合。最终选取(15, 5)组合进行后续工作。

在上述阈值条件下, 进行了试验验证。可

以看出,在聚类之前,空间数据分布特征不够明显,数据比较散乱,如图1所示。

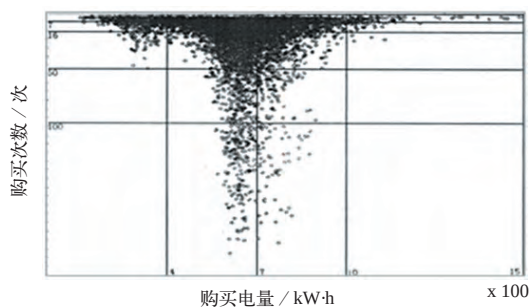


图1 聚类前数据分布图

在聚类算法之后,分布较为稀疏的数据被清除,如图2所示。从整体来看,用户数据相对集中。用电量在4~10kW·h区间范围内,购买频次为7~50。在此基础上,可以制定电力营销策略以及有效的阶梯电价方案。

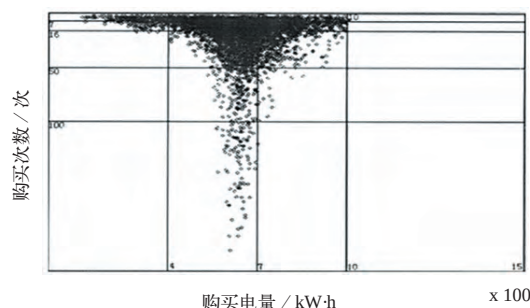


图2 聚类后数据分布图

还可以针对数据中其他维度,例如学历、孩子个数进行主成分聚类,这样可以确定学历和孩

子个数对用户用电的影响,以及基于这两个方面的用电营销策略。

5 结束语

使用聚类方法对电力营销大数据进行分析,在数据不断变化和增加的情况下,原来得到的聚类模型将难以对新数据实现匹配。后续需要进一步研究在保证聚类准确性的前提下,使聚类模式不断随着数据集的变化而优化的方法。同时,针对电力数据的展示性问题,对三维数据的可视化研究可以基于本文中的网格聚类来进行实现。

参考文献

- [1] 陈刚. 基于数据挖掘的电力营销决策支持系统的结构原理及算法研究[D]. 重庆: 重庆大学, 2004.
- [2] 李成. 数据仓库和数据挖掘在电力营销分析与辅助决策系统中的应用[D]. 南京: 东南大学, 2011.
- [3] 刘成东. 电力营销系统中数据挖掘技术的应用分析[J]. 大科技, 2014(5): 61-61, 62.
- [4] 罗静, 刘宗歧. 基于网格聚类算法的电力营销预测研究[J]. 科技信息, 2011(13): 366-367.
- [5] 罗贤缙. 聚类分析在电力营销中的应用研究[D]. 北京: 华北电力大学, 2004.

(收稿日期: 2015-05-26) EA

作者简介

慕福林(1970-), 男, 高级工程师, 主要研究方向为电力营销数据采集。

(上接第150页)

讨,分析了数据缺失的原因以及常用的补全方法。在此基础上,对朴素贝叶斯算法进行了改进,并通过三种补全方法在数据集上的实验,验证了新算法的有效性。该方法可以进一步进行电力营销数据的补全应用。

参考文献

- [1] 慕春棣, 戴剑彬, 叶俊, 等. 用于数据挖掘的贝叶斯网络[J]. 软件学报, 2000, 11(5): 660-666.
- [2] 张明卫, 王波, 张斌, 等. 基于相关系数的加权朴素贝叶斯分类算法[J]. 东北大学学报: 自然科学

版, 2008, 29(7): 952-955.

- [3] 任佳, 高晓光, 茹伟, 等. 目标数据缺失下离散动态贝叶斯网络的参数学习[J]. 系统工程与电子技术, 2011, 33(8): 1885-1890.
- [4] 廖学清, 吕强. 数据缺失下学习贝叶斯网的一种混合启发方法[J]. 计算机科学, 2008, 35(12): 163-166.

(收稿日期: 2015-05-26) EA

作者简介

慕福林(1970-), 男, 高级工程师, 主要研究方向为电力营销数据采集。