

电力营销缺失数据补全研究

针对电力数据在采集过程中的大量缺失问题进行研究,在对比一般贝叶斯方法和双尺度贝叶斯方法后,对算法进行了综合改进,使其适应于关联关系较弱、而独立性较强的电力营销数据中。同时将该算法用于数据集中进行测试,实验效果具有一定的提高。后续需要在营销系统中进行海量数据应用,使补全的数据能满足进一步营销数据挖掘的研究。

慕福林¹ 陈静² 骆凯波² 谭元刚²/1. 国网重庆电力公司 2. 国网重庆电力公司客户服务中心

0 引言

在电力营销越来越开始向智能处理方向发展的过程中,电力数据挖掘作为进行智能化的基础,一方面需要数据挖掘理论进行分析;另一方面,数据集的完整性、可用性成为新的挑战。电力营销数据来源广泛、类型复杂,在数据仓库建立过程中,数据冗余和不一致、噪声数据、表示方法错误和数据缺失等问题严重干扰了数据挖掘的结果。

本文针对数据缺失现象进行了原因分析,同时对贝叶斯方法在数据缺失补全中的应用效果进行了分析,提出双尺度贝叶斯数据补全方法,以解决数据挖掘前期数据缺失问题。

1 缺失数据

1.1 缺失数据的定义

数据缺失即数据不完整,表现在数据库中,即为该记录属性为空,通常数据缺失从缺失机制上可以分为完全性随机缺失、随机性缺失以及非完全性随机缺失。

从对象是否存在属性缺失,可以把变量分为不完全变量,即指存在数据缺失的变量或者属性;完全变量,即数据

集数据完整,不含缺失属性。

1.2 缺失数据产生的原因

现实条件下,数据缺失的主要原因可以归为以下几类。

(1) 无法有效使用的数据

这类数据,主要是指统计属性对于对象主体不具有实际意义,例如:该用户用电量为0,但其用电时间统计非0,则该数据是无效数据。

(2) 遗漏数据

例如,由于采集设备问题,数据采集过程中造成的数据缺失,同时由于采集过程不可逆,采集实时性要求很高,则不能通过其他方法获取数据。或者,由于系统设计有问题,用户在数据填写过程中,某一数据的漏填或者采集不全。

(3) 无法获取的数据

由于人为或者系统问题,在短期内无法实现采集的数据。例如,医院临床检验的时候,在一段时间内,有些数据无法得到。

(4) 获取代价很大的数据

采集需要耗费大量的人力和物力导致采集方无法进行全面的数据采集而造成的缺失。

1.3 常用的缺失数据补全方法

数据挖掘中的数据缺失,会导致挖掘结果不准确,将此结果用于实际系统,例如电力系统时,会造成极大的资源浪费,甚至会造成不可挽回的损失。因此需要对缺失数据进行补全,然后再进行数据挖掘等后续操作。

常用的数据缺失补全方法包括以下几种。

(1) 删除法

将含有缺失数据的对象全部属性删



慕福林/高级工程师

关键词/Keywords

贝叶斯·
缺失数据·
先验概率·
分布·

除, 剩余数据集作为整体进行数据处理。根据研究对象的特征, 删除法又包括配对删除、列表删除和个案删除3种。

(2) 常量补全

常量补全基于对数据集的研究上, 对于不同的数据集, 根据该对象属性的数据特性, 采用一个固定的常量值进行补全, 也可以针对数据全集进行全局数据补全。例如, 常用Null代替关系型数据库中的缺失数据。

(3) 简单补全

指利用简单易懂的公式进行数据的补全。例如, 在缺失数据附近进行欧式距离计算, 确定 K 个相邻数据, 将这几个数据进行加权均值计算, 作为补全数据。该方法的缺点是, 进行欧式距离计算时, 需要进行数据遍历, 对大量数据计算效率较低。

(4) 统计补全

通过对已知的数据集进行统计方法计算, 对缺失数据进行补全。根据数据特点, 可以对数据集进行总体统计补全, 也可以将数据集分组, 进行组内统计补全。

(5) 复杂估算方法补全

该方法是目前较为有效的缺失数据补全方法。基于全局数据, 进行模型建立, 预测每一个缺失的数据进行补全。主要方法有: 回归模型补全方法、贝叶斯补全方法、决策树补全方法和期望值最大化补全方法等。

2 贝叶斯理论在数据补全中的应用

2.1 贝叶斯理论的原理和模型

(1) 先验概率

设 S 为随机实验 E 的样本空间, 对于任意一个事件 A 有概率 $P(A)$, 且 $0 \leq P(A) \leq 1$, $P(S) = 1$ 。对于不相容事件的事件 $A_1, A_2, \dots, A_n$, 有

$$P(A) = P\left(\sum_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) \quad (1)$$

则称 $P(A)$ 为 A 发生的先验概率。由于该事件的发生不经过实验证明, 所以是通过主观判断形成的, 所以称为先验概率。

(2) 条件概率

假设 A, B 为随机实验 E 中的两个任意事件,

当 A 发的情况下, B 发生的概率, 我们表示为 $P(B|A)$

$$P(B|A) = P(AB) / P(A) \quad (2)$$

该公式为条件概率公式, $P(AB)$ 为两个事件同时发生的概率。

(3) 全概公式的定义

假设 $B_1, B_2, \dots, B_n$ 为 S 的一个划分, 且 $P(B_i) > 0, i=1, 2, \dots, n$, 则对任一 A 属于 S , 有

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n) \quad (3)$$

该公式称为全概率公式。

(4) 贝叶斯公式

假设 S 中存在 A, B 两个事件, 设 $\bar{B}$ 为 B 的逆事件, 则 A 的边际概率为

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \quad (4)$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \quad (5)$$

这是贝叶斯定理的最简单的表达式。

一般情况下, 令 $B_1, B_2, \dots, B_n$ 为样本空间 S 的一个划分, 且 $P(B_i) > 0, i=1, 2, \dots, n$, 对任意的随机事件 A 属于 S , 若 $P(A) > 0$, 则有贝叶斯公式

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^n P(A|B_i)P(B_i)}, i=1, \dots, n \quad (6)$$

2.2 贝叶斯理论的数据补全应用

在数据补全方面, 通常使用的是朴素贝叶斯分类方法。该方法具有较好的计算效率, 同时可以避免人为划分属性造成的疏漏和错误。

下面引入贝叶斯定理在数据补全方面的公式。 $C_1, \dots, C_L$ 是样本空间的划分。对样本空间的任一的事件 X , 都有

$$P(c_k | X) = \frac{P(c_k)P(X | c_k)}{\sum_{r=1}^L P(c_r)P(X | c_r)}, k=1, \dots, L \quad (7)$$

式中, $P(C_k)$ 为先验概率; $P(C_k | X)$ 为后验概率。

$P(C_k)$ 计算较为容易、对于高维数据进行 $P(X | C_k)$ 计算时, 需要的时间代价较高。因此, 朴素贝叶斯方法提出属性直接的相互独立假设, 来进行 $P(X | C_k)$ 的计算。

假设对象共有 M 个属性, 其中有一个 X_j 有类划分为 C_k , 则下面的公式成立

$$P(X|c_k) = \prod_{j=1}^{M-1} P(X_j|c_k) \quad (8)$$

应用于数据补全时, 算法描述如下:

假设数据集中的 N 条记录有 M 个属性。

N_i 代表了包含已知 X_i 属性的记录个数, N_{ik} 是当 X_i 等于它的第 K 个分类 c_{ik} 的记录个数, $N_{jr|i_k}$ 是当 X_j 等于它的第 r 个分类 c_{jr} 的记录个数, 且当 $X_i=c_{ik}, j \neq i$ 时。

属性的先验概率计算公式为

$$P(X_i=c_{ik}) = N_{ik} / N_i, i=1, \dots, L_i \quad (9)$$

当 $X_i=c_{ik}$ 时, X_j 的条件概率

$$P(X_j=c_{jr} | X_i=c_{ik}) = N_{jr|i_k} / N_{ik}, j=1, \dots, M; j \neq i; r=1, \dots, L_j \quad (10)$$

如果 X 属性缺失, 并且此属性值属于 X_i , 设 J 为 X 记录里所有非缺失值的索引集合, 计算后验概率

$$P(X_i=c_k | X) = \frac{P(X_i=c_k)}{P(X_i)} \prod_{j \in J} P(X_j=c_{jr} | X_i=c_k), k=1, \dots, L_i \quad (11)$$

3 改进的双尺度贝叶斯方法

3.1 算法概要

假设数据集中有 N 个记录, 每个记录有 M 个属性。属性设为 $X_1, \dots, X_M$, 令 $C_1, \dots, C_L$ 是某 X_i 样本空间的划分。

属性 X_i 的分类个数为 L_i , N_i 代表了包含已知 X_i 属性的记录个数。当 X_i 等于它的第 K 个分类时, c_{ik} 的记录个数设为 N_{ik} 。 $N_{jr|ik}$ 是当 X_j 等于它的第 r 个分类 c_{jr} 的记录个数。

计算每一个属性的先验概率

$$P(X_i=c_k) = N_{ik} / N_i \quad (12)$$

$i=1, \dots, M; k=1, \dots, L_i$

计算当 $X_i=c_{ik}$ 时, 属性 X_j 的后验概率

$$P(X_j=c_{jr} | X_i=c_{ik}) = N_{jr|ik} / N_{ik} \quad (13)$$

$j=1, \dots, M; j \neq i; r=1, \dots, L_j$

当记录 x 的属性 X_j 存在一个缺失时, 令 J 为 x 的非缺失数据索引。当 $X_i=c_{ik}$ 时, 计算 X_j 条件概率, 修正因子计算公式为

$$\theta = P(X_j | X_i=c_{ik}) = N_{jr|ik} / N_{ik}, \quad (14)$$

$$X_j = X_1 \cap X_2 \cap \dots \cap X_j \cap \dots \cap X_M; j \neq i$$

后验概率 P_1 的计算公式为

$$P_1 = P(X_i=c_k | X_j) = \frac{1}{P(X_i)} P(X_i=c_k) \prod_{j \in J} P(X_j=c_{jr} | X_i=c_k), k=1, \dots, L_i \quad (15)$$

后验概率 P_2 的计算公式如下

$$P_2 = P(X_i=c_k | X_j) = \frac{1}{P(X_i)} P(X_i=c_k) \min(P(X_j=c_{jr} | X_i=c_k)), k=1, \dots, L_i \quad (16)$$

最终计算公式为

$$P(c_k / X)' = P_1 + (P_2 - P_1)\theta \quad (17)$$

3.2 算法分析

假设属性的平均分类为 L , 数据总量为 N , 则算法时间复杂度为 $O(NML)$ 。当 N 足够大时, 时间复杂度与 N 呈线性正比关系。

空间复杂度与计算过程中存储的中间值相关, 一般来说, 后验概率值不需要存储, 缺失填充后就删掉进行下一轮计算。则先验和条件概率分别是 $O(ML)$ 和 $O(M^2L^2)$ 。

此外, 该算法对数据集和参数不做具体设定, 具有普遍适用性。

3.3 数据填充

当计算出缺失数据概率后, 可以采用不同的策略进行缺失数据填充。

(1) 最大概率填充法

即使用概率最大的分类值来进行缺失数据填充。

顾名思义, 就是使用最大概率的分类值来填充缺失数据。

对样本数据进行双尺度贝叶斯缺失数据预测, 样本数据库如表1所示。

表1 样本数据库

No.	Income	Age	Gender	HomeOwner
1	low	<30	female	no
2	low	<30	male	no
3	low	30-55	female	yes
4	low	30-55	female	no
5	low	>55	female	no
6	high	<30	male	yes
7	high	30-55	female	yes
8	high	30-55	male	yes
9	high	30-55	male	yes
10	high	30-55	male	no
11	high	>55	male	yes
12	?	30-55	female	yes
13	?	30-55	female	yes
14	?	<30	female	?
15	?	?	male	no

表1第12条中Incom缺失数据high的预测结果大于low, 则对其进行high填充。同样的方法用于其他缺失数据。则表格补全后, 如表2所示。

表2 MaxPost填充缺失数据

No.	Income	Age	Gender	HomeOwner
12	high	—	—	—
13	high	—	—	—
14	low	—	—	no
15	high	<30	—	—

(2) 概率分布 (PropPost) 填充法

针对表1里的第12条和第13条缺失数据来说, Income的“low”和“high”的预测概率分别为0.420 1和0.579 9。因此, 在这两条记录里面的“Income”值, 更可能有一个“low”, 另一个是“high”。填充缺失数据后如表3所示。

表3 PropPost填充缺失数据

No.	Income	Age	Gender	HomeOwner
12	high	—	—	—
13	low	—	—	—
14	low	—	—	no
15	high	<30	—	—

(3) 权值填充法

将缺失数据的概率参数填充到数据集中, 补全结果如表4所示。

表4 带权值的方法填充缺失数据

No.	Income	Age	Gender	HomeOwner
12	high(0.63)	—	—	—
13	high(0.63)	—	—	—
14	low(0.98)	—	—	no(0.68)
15	high(0.49)	<30(0.53)	—	—

3.4 算法实验及结果分析

本实验选取了机器学习仓库中的3个数据集进行验证。第一个是病例数据集Breast Cancer, 具体数据包括: 结果、身体状况、年龄和测试方法等9个分类属性; 第二个是车辆数据集Car Evaluation, 共包含6个分类属性, 每个属性至少有4个子分类, 1 700条记录; 第三个数据集为Nursery, 共12 960条记录, 8个分类属性。

由于实验为缺失数据补全验证, 将数据集中的

数据进行随机缺失设置。通过预测后的填充结果与之前数据进行比较, 来验证算法的有效性, 三种方法预测的准确率如表5所示。

准确率 = (缺失数据中预测正确的个数/缺失数据的总个数) × 100%

表5 三种方法预测的准确率

补全方法名称		Breast Cancer (%)	Car Evaluation (%)	Nursery
最高频率 补全	MaxFreq	33.87	48.01	50.39
朴素贝叶斯 补全	MaxPost	43.62	50.22	53.54
双尺度贝叶斯 补全	MaxPost	44.51	52.34	53.96

实验表明, 最高频率补全方法的效果最差, 双尺度贝叶斯效果最好。因为双尺度贝叶斯对朴素贝叶斯进行了改进, 同时使用最大概率补全。

错误率 E 的计算公式为

$$E = \frac{1}{|T|} \sum_{t=1}^{|T|} \frac{|Q_t - Q'_t|}{Q_t} \quad (18)$$

式中, T 为所有的频繁项集; Q_t 是指原数据中第 t 个频繁项集的支持度; Q'_t 为经过人为删除25%的数据后又补全的数据库中相同的频繁项集的支持度。

试验结果如表6所示。

表6 错误率 E 对比结果

补全方法名称		Breast Cancer	Car Evaluation	Nursery
最高频率补 全法	MaxFreq	40.07	45.09	41.38
朴素贝叶 斯补全法	MaxPost	17.98	11.69	15.00
双尺度贝叶 斯补全法	PropPost	9.85	5.48	3.75
双尺度贝叶 斯补全法	MaxPost	10.65	7.47	11.73
双尺度贝叶 斯补全法	PropPost	5.23	4.25	3.41

实验结果表明: 双尺度贝叶斯方法具有较高的补全效果, 最大概率补全的效果, 远低于分布补全方法。

4 结束语

本文通过对数据挖掘中, 数据缺失问题进行探
(下转第154页)

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。

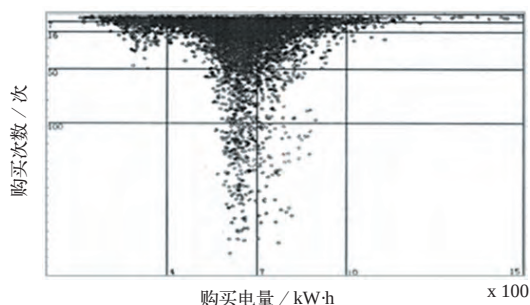


图1 聚类前数据分布图

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。

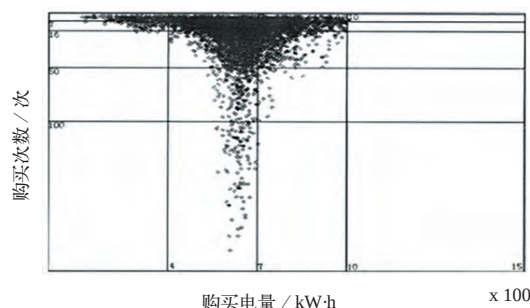


图2 聚类后数据分布图

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。