

电动汽车运营管理系统数据质量分析及优化策略

电动汽车充电服务的数据质量问题对运营管理系统运行和大数据分析技术应用都至关重要。针对北京地区电动汽车智能充换电运营管理系统的充电交易记录数据，通过分析异常数据的类型和分布情况，基于业务数据之间的数理和逻辑关系，提出异常数据的筛选规则和处理方法，给出了系统优化策略方案。通过在系统中设计与应用数据筛选与处理模块，能够及时发现并修正异常数据，指导系统和设备运维，有效提高数据质量，对比分析结果验证了所提方案的可行性和有效性，为大数据分析技术的实施应用提供保障。

张禄¹ 潘鸣宇¹ 陈艳霞¹ 迟忠君¹ 赵越² 刘诗²/1.国网北京市电力公司电力科学研究院 2.北京瑞天科技有限公司



张禄/工程师

关键词/Keywords

电动汽车 ·
运营管理系统 ·
大数据 ·
数据质量 ·
数据筛选与处理 ·

0 引言

在能源短缺、环境污染严重的背景下，作为一种新型交通工具，电动汽车凭借其无污染或低污染物排放、低噪声、高效、维修及运行成本低等优点，成为当代交通工具发展的主流方向，也成为各国政府、汽车制造商和能源企业关注的焦点^[1]。

北京是国内首批推动电动汽车发展和示范运营的城市^[2]。近年来，北京地区电动汽车数量迅速增长，截止2014年底拥有各类电动汽车合计13374台，其中私人乘用车4050台。电动汽车智能充换电运营管理系统（简称运营管理系统）对北京地区200余座充电站的充电桩进行信息化运营和服务管理，为电动公交车、出租车、环卫车以及私人乘用车提供充换电服务^[3]，为合理规划充电设施并提升用户充电体验，充分挖掘利用运营管理系统积累的大量运营数据，基于运营

管理系统建立了数据分析平台，能够支持电动汽车运营监控和大数据分析技术应用。电动汽车大数据分析平台系统结构如图1所示。

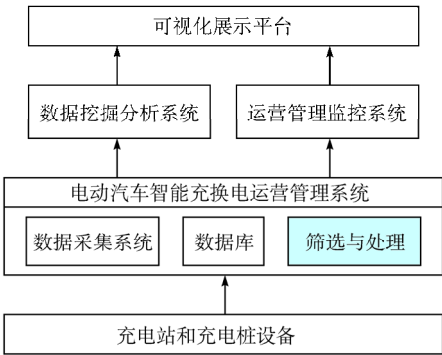


图1 电动汽车大数据分析平台系统结构

本文首先介绍了数据筛选与处理技术（ETL，抽取、转换和加载），对运营管理系统的数据质量和异常现象进行分析，给出了分析结果，提出了系统优化方案。通过在实际运营管理系统中加入数据筛选与处理模块，提高了分析数据的准确率和可用性，为业务部门持续改进数据采集质量提供有效手段，对比结果验证了所提优化方案的有效性。

1 数据筛选

1.1 数据筛选的必要性

运营管理系统数据库存储的原始数据受到终端设备采集准确度、通信信道干扰和系统设计缺陷等因素的影响，通常需要经过数据筛选和处理环节优化数据质量，在众多信息中提取出有效的数据项作为分析模型的输入变量。数据筛选的必要性包括：

1) 提高分析模型的稳定性。过多

的输入变量可能带来干扰和过拟合等问题,精简而准确的输入变量可以保证模型具有较好的稳定性。

2) 提高分析模型预测能力。过多的输入变量会存在共线性问题,增加对模型计算结果的解释难度,容易引起模型参数震荡,降低模型的预测能力。

3) 提高分析系统运算速度和效率。

1.2 数据筛选的方法

数据筛选的思路主要包括 3 个方面:结合业务经验的数理和逻辑关系进行先行筛选;采用线性相关性指标对输入自变量进行初步筛选;针对分析目标变量对输入自变量进行相关性判断和选择^[4]。在数据质量治理初期主要采取第一种筛选方法。

1.2.1 结合业务经验的先行数据筛选

结合业务经验对原始数据进行筛选往往是最重要且有效的方法。通过梳理数据在实际业务层面上的内在含义和关系,借助专家经验和判断,确定数据筛选的数理或逻辑关系,从而发现异常数据并设置筛选规则修正和剔除问题数据。

在运营管理系统中,针对电动汽车充电交易记录及其异常数据特点,设置了 5 项异常数据筛选规则,并对部分异常数据进行了修复。

1.2.2 基于线性相关性的输入变量间数据筛选

对分析模型的输入自变量进行筛选的最简单且常用的方法就是采用线性相关指标,其中以皮尔逊相关系数(Pearson Correlation)最为常用。Pearson 相关系数主要用于比例型变量之间、区间型变量之间以及二元变量之间的线性关系描述^[5]。其计算公式为

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \tag{1}$$

式中,分子部分为 x 与 y 的协方差;分母部分为 x 标准差与 y 标准差的乘积;相关系数 r 的取值范围为 $[-1,+1]$, r 的大小表示不同程度的线性相关程度; $|r|<0.3$,表示低度线性相关; $0.3 \leq |r| < 0.5$,表示中低度线性相关; $0.5 \leq |r| < 0.8$,表示中度线性相关; $0.8 \leq |r| < 1.0$,表示高度线性相关。

在建模时的变量筛选过程中,对于输入变量之间属于中度以上线性相关的(>0.6 以上),可

以选择保留其中一个。

Pearson 相关系数指标除了线性相关之外,并不能排除变量之间存在其他形式的非线性相关关系,这也是该指标筛选的局限性。

1.2.3 针对目标变量相关性的输入变量数据筛选

R^2 二次方(也称作 R^2 或 Coefficient of Multiple Determination)方法,采用多元线性回归的分析算法判断和选择对目标变量有重要影响和预测意义及价值的输入变量。 R^2 表示模型输入的各变量在多大程度上可以解释目标变量的可变性,其取值范围在 $[0,1]$ 之间, R^2 越大,说明模型的拟合度越好。 R^2 的计算公式为

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{2}$$

式中, y_i 为目标变量的真实值; f_i 为模型的预测值; $\bar{y}$ 为目标变量真实值的均值。

R^2 方法更具针对性地考虑分析目标与输入变量之间的相关关系,从而有效筛选出最有分析价值的

2 系统异常数据分析

2.1 系统数据情况

北京地区充电站主要包括公交车充换电站、环卫车充换电站、出租车充电站、乘用车充电站以及物流车充电站五类站点,充电桩包括慢速交流充电桩和快速直流充电桩,充电交易记录包括用户一次充电的时间、电量和电费等信息。截至 2014 年底,电动汽车运营管理系统积累的充电设施资产数据和充电交易记录情况如表 1 所示。

表 1 系统数据情况

数据项名称	数据项含义	数据量
CHARGING STATION INFO	充电站信息	232
CHARGING PILE INFO	充电桩信息	1 986
TRANSACTION RECORDS	充电交易记录	1 008 539

将充电站、充电桩资产信息与充电交易记录数据进行匹配合并,得到包含对应充电站、桩信息的充电交易记录 1 008 539 条。对所有交易记录进行探索性数据分析(EDA 分析),发现在筛

选剔除疑似异常数据后,最终剩余的正常数据量为 657 104 条,系统数据的平均可用率为 65.15%,其中系统运行初期的异常数据比重较大。

2.2 分析变量定义

为方便分析和表述,结合业务数据含义和关系,充电交易记录中的变量定义如表 2 所示。

表 2 分析变量定义

变量名	变量定义	变量单位
T_D	充电时长	min.
F	充电电费	RMB
P	充电电价	RMB/kW·h
$E(E_{sta}; E_{end})$	充电量(充电前;充电后)	kW·h
$B(B_{sta}; B_{end})$	充电卡金额(充电前;充电后)	RMB

用户一次充电的电量由充电前后充电桩内电能表的表底数计算得到,计算公式为

$$E = E_{end} - E_{sta} \tag{3}$$

用户一次充电的金额由充电前后用户电卡内余额计算得到,计算公式为

$$B = B_{sta} - B_{end} \tag{4}$$

充电电费由充电电量和电价计算得到,用户一次充电的电费要与充电金额相等,计算公式为

$$F = E \times P \tag{5}$$

2.3 异常数据分析

为了精准掌握异常数据情况,对上述各类异常数据的时间和空间分布进行详细分析,并结合业务经验给出了导致数据异常的原因。

2.3.1 交易记录数据项与相关资产信息不符

此类异常数据指在充电交易记录中缺少相匹配的电表、充电桩 ID 或充电站 ID 信息的情况,主要集中于系统运营早期阶段。其原因主要包括:临时增加的充电测试点没有在系统档案中登记;充电设备运维商在更新电表后未及时上报系统导致资产匹配失败。

2.3.2 充电交易日期异常

充电交易日期和时间由充电设备提供。充电桩设备除了自身带有时钟计时之外,还要每天零点完成与系统的自动对时,保证设备时钟的准确性。

分析可知,该类异常数据主要包括三种情况:充电日期明显超出合理范围、充电开始时刻晚于结束时刻以及充电时长明显超出合理范围。充电日期异常主要集中在 3 个出租车充电站。

造成此类异常数据的主要原因包括:充电桩与系统之间的自动对时功能存在问题;系统在采集数据时未对充电日期准确性进行校验。

2.3.3 充电交易电量、电费及金额异常

用户充电交易中的充电量由充电前后电能表的表底值之差计算得到,充电电费由电量乘以电价得到。此外充电金额由充电前后用户充电卡内余额之差计算得到,充电金额须等于充电电费。

分析可知,充电电费不等于充电量与电价乘积值的情况主要发生在 2014 年 10、11 月份的 A 区和 B 区公交车充电站。业务分析判断其原因是该站点大功率充电桩用于计算电费的电量值未计及电流互感器的变比系数,导致少计电费和少扣金额。

2.3.4 充电时长异常

充电时长由充电开始时刻与结束时刻计算得到。用户一次充电交易记录中,充电时长为 0 而充电量不为 0,这种异常数据主要集中在 2013 年 12 月份以及 2014 年 1 月份,且主要发生在 E 区等 5 个出租车充电站。因为充电开始时刻等于结束时刻将导致充电时长为 0,所以充电桩设备时钟故障会导致一次正常的充电行为记录下来的交易记录中电量不为 0 而时长为 0。可见,设备时钟问题是造成该类异常数据的主要原因。

2.3.5 充电量异常

用户一次充电交易记录中,充电时长不为 0 而充电量为 0,这种异常数据主要集中在 2014 年 1、2 和 7 月份,且主要发生在 I 区等 5 个出租车充电站。考虑到该类异常数据有可能在实际充电操作中发生,所以该类数据作为特殊的用户行为进行分析。

2.3.6 交易记录出现重复

交易记录出现重复的数据主要集中在 2014 年 1 月,且主要发生在 4 个出租车充电站。数据重复的原因是系统数据信道问题造成充电桩向系统重复推送交易记录,而采集系统未对上传的重复数据进行筛选。

3 数据系统优化

3.1 优化方案

通过系统异常数据的分析结果可见，初期数据采集系统仅被动接收数据，没有对数据进行合理的筛选与处理。因此，为了保证分析数据的质量，需在数据采集和存储过程中增加数据筛选与处理模块。结合实际异常数据类型，提出了数据采集系统的优化改造方案，其中数据筛选与处理流程如图 2 所示。

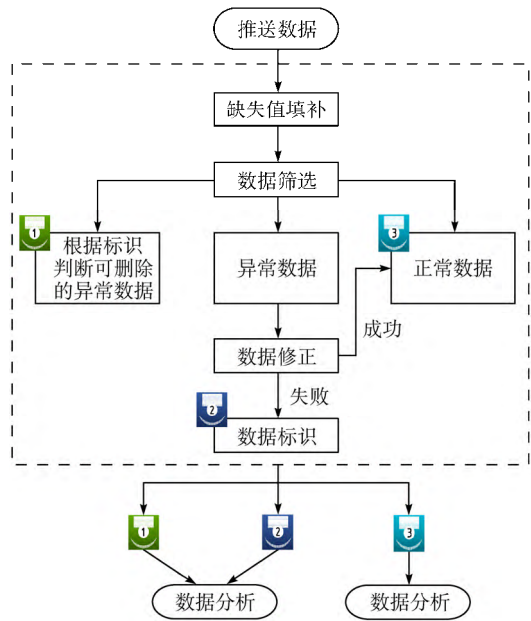


图 2 数据采集系统中数据筛选与处理流程图

数据筛选与处理模块各部分含义如下：

缺失值填补模块是指通过充电卡号、结算设备 ID 等主要字段进行联合匹配，对缺失充电站/桩信息进行填补；对无法修正的缺失数据添加缺失标识。

数据筛选模块指根据数据筛选规则筛选出正常数据、重复数据和异常数据。

运营管理系统针对电动汽车充电交易记录异常数据类型特点，设置异常数据筛选规则如下：

- 1) 规则 1。充电交易时间异常，包括不在正常范围内、交易时间晚于记录载入系统时间。
- 2) 规则 2。充电量异常，交易记录中用户一次充电量不等于充电桩电表表底前后值之差，即 $E \neq E_{\text{end}} - E_{\text{sta}}$ 。

3) 规则 3。充电金额异常，交易记录中用户一次充电电费不等于充电卡前后余额之差，即 $F \neq B_{\text{sta}} - B_{\text{end}}$ 。

4) 规则 4。电费计算准确度，用户一次充电电费不等于充电量乘以充电电价，即 $F \neq E \times P$ 。

5) 规则 5。充电时间和电量合理性，用户一次充电时间为零而充电量不为零，或者充电量为零而充电时间不为零。

筛查出来的异常数据可概括分为 3 种类型：

1) 第一类异常数据。经过有效处理修复后转换成正常数据重新参与分析，包括通过上述规则 1、2、3 和 4 筛选出来的部分异常数据。

2) 第二类异常数据。与设备缺陷或用户行为等因素密切相关，具有较高的分析价值，包括通过上述规则 1、4 和 5 筛选出来的部分异常数据。

3) 第三类异常数据。无法修复再利用的数据，只能排除在分析工作之外，包括通过上述规则 1 筛选出来的部分异常数据。

(1) 数据修正

数据修正指通过校验交易记录录入时间，进而对充电时间错误数据进行修正，对无法进行校正的时间错误数据添加错误标识。

(2) 异常数据标识

为方便做进一步的筛选和修复处理，对异常数据进行标识的规则如表 3 所示。

表 3 异常数据标识

标识类型	规则定义
ERROR_TYPE=1	$F \neq B_{\text{sta}} - B_{\text{end}}$
ERROR_TYPE=2	$F \neq E \times P$
ERROR_TYPE=3	$E \neq E_{\text{end}} - E_{\text{sta}}$
ERROR_TYPE=4	$T_D=0$ 且 $E>0$
ERROR_TYPE=5	$E=0$ 且 $T_D>0$

通过在数据采集系统中加入数据筛选与处理模块，系统依次完成数据的接收、筛查和异常数据的修正，从而保证数据处理的连贯性。若数据无法修正，则为该数据添加标识并在第一时间通知运维人员对其进行确认，及时完成消缺处理，支持现场运维人员提高设备检修效率，从而进一步保证数据质量。

3.2 优化成效验证

为了验证提出的采集系统优化方案的应用效果,通过在数据采集系统中加入数据筛选与处理模块,实现了上述数据筛查和修正功能。对系统实现优化后 2 个月的共计 155 991 条数据进行分析,可得到以下结论:

- 1) 可用于后续分析的正常数据条数为 137 116,即数据的准确度为 87.90%,与之前没使用数据处理模块相比,准确度提升了 22.75%。
- 2) 经过数据处理模块处理后的数据,所有类型的错误数据均有所减少:部分日期异常数据得到修正,修复数据 3.87%;经过运维人员现场消缺排查,充电交易金额、电量及电费异常数据被消除,其他异常数据类型均有所下降。
- 3) 系统数据处理优化方案的实施,在筛选和修正异常数据、指导现场运维方面都取得了显著成效。

4 结束语

本文针对电动汽车智能充换电运营管理系统的充电交易记录数据质量问题,分析了系统初期存在的异常数据类型和充电站分布情况。基于业务数据具有的数理和逻辑关系,提出了异常数据筛选规则以及修正和标识处理方法,给出了数据采集系统数据质量优化方案。通过在系统中增加数据筛选与处理模块,有效降低了系统异常数据比例,改善后的数据质量优化成效对比结果验证了所提方案的可行性和有效性。

通过运营管理系统数据分析的相关研究和实践发现,提高数据质量一方面需要系统的数据

筛选与处理技术;另一方面还要提高现场设备的数据采集质量和运维管理水平。数据质量分析发现系统和设备问题并给出解决方案,为充电设施运营提供有力支持,并为大数据分析工作打下坚实基础。

参考文献

[1] SONG Y H, YANG X, LU Z X, et al. Integration of plug-in hybrid and electric vehicles :experience from China [C]. Minneapolis : Proceedings of Power & Engineering Society General Meeting, 2010, July 25-29, 1-5.

[2] 陈科,王震坡,林程.北京市纯电动汽车的发展与示范运营[J].新材料产业,2006,9:27-30.

[3] 潘鸣宇,贺家胜,迟忠君,等.电动汽车智能服务平台的设计与实现[J].电力信息与通信技术,2014,12(1):78-81.

[4] 卢辉.数据挖掘与数据化运营实战:思路、方法、技巧与应用[M].北京:机械工业出版社,2015.

[5] 王涓,吴旭鸣,王爱凤.应用皮尔逊相关系数算法查找异常电能表用户[J].电力需求侧管理,2014,16(2):52-54.

作者简介

张禄(1984-),男,博士,工程师,研究方向为电动汽车运营服务及大数据分析技术。

潘鸣宇(1985-),男,高级工程师,研究方向为电动汽车运营管理及信息化技术。

(收稿日期:2015-12-10) EA
(编辑:史海疆)

(上接第 21 页)

核的种类已经包含了如设备名称校核、状态校核、潮流校核和五防校核等一系列的校核方式,而由于校核所系统的接口是统一的,随时可以按照接口规范新增其他的校核方式。

2 结束语

在电网运作过程中,口令操作校核是一个空白区域。本文针对口令操作,研究口令操作的流程,并对该流程中的每一个流程点通过对现有的成熟技术进行研究和借鉴提出实现方式,最终提

出并设计了一种基于语音转换校核的口令操作安全校核系统,并初步进行了可行性的探讨。但是本系统在实际中的应用成果还需要进行验证,各个流程节点的技术实现也需要不断进行完善和改进。

作者简介

叶海峰(1978-),男,硕士,高级工程师,主要从事工作领域:电网运行。

(收稿日期:2015-12-10) EA
(编辑:史海疆)