(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110962684 A
(43)申请公布日 2020.04.07

(21)申请号 201911117375.5

(22)申请日 2019.11.15

(71)申请人 东华大学
地址 201600 上海市松江区人民北路2999号

(72)发明人 张光林 黄淦 赵萍

(74)专利代理机构 上海申汇专利代理有限公司 31001
代理人 徐俊 柏子雯

(51)Int.Cl.
B60L 58/10(2019.01)
B60L 1/00(2006.01)
B60L 58/24(2019.01)

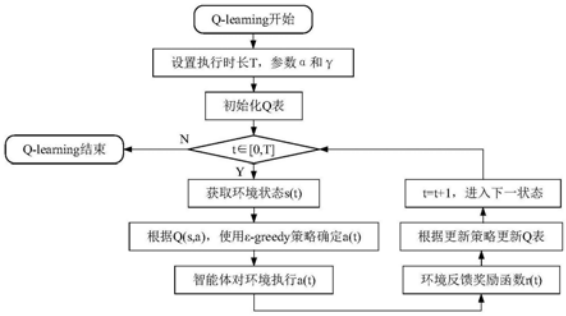
权利要求书2页 说明书6页 附图2页

(54)发明名称

电动汽车能源管理与分配方法

(57)摘要

本发明涉及一种电动汽车能源管理与分配方法，一个实施例的方法包括：建立电动汽车的系统模型，对汽车不同的工况进行量化，确定能源管理系统的状态空间；确定优化目标函数，使得电动汽车在一个行驶周期内的耗能最少；使用强化学习的方法对目标函数进行优化，确定马尔可夫决策过程的状态转移概率及回报函数；在多次工况运行后生成“状态-动作映射”Q矩阵，并对其不断进行更新，获得最佳的能源管理策略。本实施例方案提高了电动汽车在行驶过程中的能源利用效率。



CN 110962684 A

1.一种电动汽车能源管理与分配方法，其特征在于，包括以下步骤：

步骤1、将电动汽车的电力设备分为高压供电网网络、低压供电网络、充电网络三大类，其中：高压供电网络包含热管理系统和驱动系统；低压供电网络用以支持全车低压电子部件的用电；充电系统用于从电网向车载电池补充能量，建立电动汽车能源系统模型如下式(1)所示，根据现实中复杂多变的汽车运行工况，对汽车不同的工况进行量化，确定能源系统的状态空间：

$$\begin{cases} \min\{ECR|k_{AC}(t)\}, t \in (0,T) \\ P_{mot}(t) = P_{HV}(t) - k_{AC}(t)P_{AC-max}(t) \\ P_{mot}(t) \geq P_{drv}(t), t \in (0,T) \end{cases} \quad (1)$$

式(1)中，ECR表示能源消耗率；$k_{AC}(t)$表示热管理系统功率系数，表示将高压供电系统分配给热管理系统的功率占热管理系统峰值功率的比例；T表示一个行驶周期时长；$P_{mot}(t)$表示向驱动系统输出的最大功率；$P_{HV}(t)$表示可被能源管理系统所分配的高压供电网络的总功率；$P_{AC-max}(t)$表示电动汽车的制冷设备的峰值功率；$P_{drv}(t)$表示电动汽车的驾驶员所期望的驱动系统实际输出功率；

步骤2、根据步骤1建立的电动汽车能源系统模型确定优化目标函数，使得电动汽车在一个行驶周期内的耗能最少，将求解ECR最小化的问题转化为求解0～T时间段内电动汽车在[0,T]时间段内总电量消耗E(t)最小化的问题，E(t)如下式(2)所示：

$$\begin{aligned} E(t) &= E_{mot}(t) + E_{AC}(t) + E_{LV}(t) + E_{loss}(\Delta T(t)) \quad (2) \\ &= \int_0^t [P_{mot}(\tau) + k_{AC}(\tau) \cdot P_{AC-max} + P_{LV}(\tau)]\, d\tau + E_{loss}(\Delta T(t)) \end{aligned}$$

式（2）中，$E_{mot}(t) = \int_0^t P_{mot}(\tau)d\tau$是驱动系统所消耗总能量；$E_{AC}(t) = \int_0^t k_{AC}(\tau) \cdot P_{AC-max}d\tau$是热管理系统所消耗的能量；$E_{LV}(t) = \int_0^t P_{LV}(\tau)d\tau$是车辆低压供电网络所消耗的能量，$P_{LV}(\tau)$表示低压供电网络所需的总功率；$E_{loss}(\Delta T(t))$是0～t时间段内电池组温度变化而耗散的能量；

步骤3、使用强化学习的方法对步骤2确定的目标函数进行优化，强化学习模型采用Markov决策过程；

在车辆运行过程中，使用Q-learning算法对能量分配策略进行学习和优化，Q-learning算法其Q矩阵采用以下的策略不断更新，以得到更为节能的能源分配策略使得车辆的能源管理系统拥有一个状态动作值矩阵，即Q表，Q表内的每一项Q(s,a)为状态s与动作a的映射关系，当电动汽车能源管理系统在某一状态s下进行能量分配时，Q表采用式(3)的方法进行更新：

$$Q(s,a)_{t+1} \leftarrow Q(s,a)_t + \alpha[r_t + \gamma \max_{a(t+1)} Q(s,a)_{t+1} - Q(s,a)_t] \quad (3)$$

式(3)中，$Q(s,a)_t$是在采取动作a之前智能体对映射(s,a)的估计；$r_t + \gamma \max_{a(t+1)} Q(s,a)_{t+1}$是Q(s,a)的现实值；α是学习率，表示Q值的过去值和新获得奖励的加权关系；$\gamma \in [0,1]$是衰减系数，反应未来奖励对当前决策的重要性，γ越大表示智能体在采取动作a是越倾向于考虑未来奖励的影响；

步骤4、当$Q(s,a)_t$收敛于最优映射时，认为系统已经完成了学习过程，即Q-learning算

法得到的最优策略为：

$$a^* = \pi(s) = arg\,\max_a Q(s,a) \qquad (4)$$

式(4)中，arg max是对函数求参数的函数，即求得使得策略π(s)最大的Q表映射，对于离散性的马尔可夫决策问题，这是使得系统策略最佳的必要条件，即对于环境所处的任意状态s，能量管理系统总能选择使得Q值最大的动作a^*。

2.如权利要求1所述的一种电动汽车能源管理与分配方法，其特征在于，将新欧洲标准循环测试NEDC的循环测试工况表示为只有加速、减速、等速和停车四种状态；通过电动汽车的参数表计算出车辆处于每种测试状态时的电动机输出功率$P_{mot}(t)$。

3.如权利要求1所述的一种电动汽车能源管理与分配方法，其特征在于，将驱动系统的功率$P_{mot}(t)$和电池组的温度ΔT(t)定义为步骤3中强化学习模型的状态空间，记为$s(t) \triangleq \left(P_{mot}(t), \Delta T(t)\right)$；

将电动汽车的制冷系统的运行功率分配系数$k_{AC}(t)$定义为步骤3中强化学习模型的动作空间；

步骤3中强化学习模型的奖励函数定义为$r(t)=r_{avg}(t)-E(t)$，其中，$r_{avg}(t)$为0～t时刻智能体所获得的奖励的平均值，对于时间离散的系统奖励的平均值$r_{avg}(t_n)=\frac{1}{n}[r(t_1)+r(t_2)+\cdots+r(t_n)]$，$r(t_n)$为$t_n$时刻智能体所获得的奖励。

4.如权利要求1所述的一种电动汽车能源管理与分配方法，其特征在于，在所述Q-learning算法的基础上引入ε-greedy策略，使得系统不仅会采取由Q表中所得出的最优动作，并且会以某个概率对当前Q值并非最大的动作进行试探。

电动汽车能源管理与分配方法

技术领域

[0001] 本发明涉及信息处理技术领域，特别是涉及一种电动汽车能源管理与分配策 略以及其实现算法。

背景技术

[0002] 目前，推动汽车工业的节能化、信息化发展是大势所趋，大力推进传统汽车 工业向新能源汽车转型升级也已成为全球汽车产业的首要任务。电动汽车包括纯 电动汽车、混合动力汽车以及燃料电池汽车等。就目前的发展状况而言：混合动 力汽车可以在一定程度上缓解汽车对石油能源的依赖，然而其终究无法实现无污 染和零排放，只是一种向新能源过渡的暂时性方案；纯电动汽车兼具能源效率高、 零排放、噪声小、维修方便、结构简单等优点，而且还可使用非化石燃料的其他 能源转化为电能，是理想的实现节能减排的选择。

[0003] 然而，由于电池储能技术长时间未取得革命性突破，续航里程短、充电时间 长的问题长期以来制约着电动汽车的大规模推广和应用。在电动汽车能源管理领 域，基于规则的能量管理策略在实际应用中最为广泛，但这种策略不具有动态优 化的特性，它们无法在复杂的行驶环境下充分发挥电动汽车节能潜质。还有一种 基于优化的能源管理策略，这种策略需要将待优化的问题转化为若干数学约束式 进行表示，并确定求解目标、设计反馈函数，利用优化算法探寻达到最值的方法， 其缺点是需要预知行驶工况、计算量大、灵活性差，无法保证全局最优，使其实 际优化效果大打折扣。有研究人员将机器学习应用于混动汽车的能源管理，通过 不断尝试获得赏罚信息的方式进行探索，脱离了对被控系统模型的依赖。但是学 习的方法有赖于执行完一个动作之后环境所反馈的奖励信号，这种反馈总是存在 不可避免的噪声和延时。

发明内容

[0004] 本发明的目的是：降低电动汽车在行驶过程中的能量损耗，提高电动汽车的 能源利用率。

[0005] 为了达到上述目的，本发明的技术方案是提供了一种电动汽车能源管理与分 配方法，其特征在于，包括以下步骤：

[0006] 步骤1、将电动汽车的电力设备分为高压供电网网络、低压供电网络、充电 网络三大类，其中：高压供电网络包含热管理系统和驱动系统；低压供电网络用 以支持全车低压电子部件的用电；充电系统用于从电网向车载电池补充能量，建 立电动汽车能源系统模型如下式(1)所示，根据现实中复杂多变的汽车运行工 况，对汽车不同的工况进行量化，确定能源系统的状态空间：

[0007]
$$\begin{cases} \min\{ECR|k_{AC}(t)\}, t \in (0,T) \\ P_{mot}(t) = P_{HV}(t) - k_{AC}(t)P_{AC-max}(t) \\ P_{mot}(t) \geq P_{drv}(t), t \in (0,T) \end{cases} \quad (1)$$

[0008] 式(1)中，ECR表示能源消耗率；$k_{AC}(t)$表示热管理系统功率系数，表示 将高压供电

系统分配给热管理系统的功率占热管理系统峰值功率的比例;T表示 一个行驶周期时长;$P_{mot}(t)$表示向驱动系统输出的最大功率;$P_{HV}(t)$表示可被能 源管理系统所分配的高压供电网络的总功率;$P_{AC-max}(t)$表示电动汽车的制冷 设备的峰值功率;$P_{drv}(t)$表示电动汽车的驾驶员所期望的驱动系统实际输出功 率;

[0009] 步骤2、根据步骤1建立的电动汽车能源系统模型确定优化目标函数,使得 电动汽车在一个行驶周期内的耗能最少,将求解ECR最小化的问题转化为求解 0～T时间段内电动汽车在[0,T]时间段内总电量消耗E(t)最小化的问题,E(t)如下 式(2)所示:

[0010]
$$E(t) = E_{mot}(t) + E_{AC}(t) + E_{LV}(t) + E_{loss}(\Delta T(t)) \quad (2)$$
$$= \int_0^t [P_{mot}(\tau) + k_{AC}(\tau) \cdot P_{AC-max} + P_{LV}(\tau)]\, d\tau + E_{loss}(\Delta T(t))$$

[0011] 式(2)中,$E_{mot}(t) = \int_0^t P_{mot}(\tau)d\tau$是驱动系统所消耗总能量;$E_{AC}(t) = \int_0^t k_{AC}(\tau) \cdot P_{AC-max}d\tau$是热管理系统所消耗的能量;$E_{LV}(t) = \int_0^t P_{LV}(\tau)d\tau$是车辆低 压供电网络所消耗的能量,$P_{LV}(\tau)$表示低压供电网络所需的总功率;$E_{loss}(\Delta T(t))$ 是0～t时间段内电池组温度变化而耗散的能量;

[0012] 步骤3、使用强化学习的方法对步骤2确定的目标函数进行优化,强化学习 模型采用Markov决策过程;

[0013] 在车辆运行过程中,使用Q-learning算法对能量分配策略进行学习和优化,Q-learning算法其Q矩阵采用以下的策略不断更新,以得到更为节能的能源分配 策略使得车辆的能源管理系统拥有一个状态动作值矩阵,即Q表,Q表内的每一 项Q(s,a)为状态s与动作a的映射关系,当电动汽车能源管理系统在某一状态s下 进行能量分配时,Q表采用式(3)的方法进行更新:

[0014] $Q(s,a)_{t+1} \leftarrow Q(s,a)_t + \alpha[r_t + \gamma \max_{a(t+1)} Q(s,a)_{t+1} - Q(s,a)_t]$ (3)

[0015] 式(3)中,$Q(s,a)_t$是在采取动作a之前智能体对映射(s,a)的估计;$r_t + \gamma \max_{a(t+1)} Q(s,a)_{t+1}$是Q(s,a)的现实值;α是学习率,表示Q值的过去值和 新获得奖励的加权关系;$\gamma \in [0,1]$是衰减系数,反应未来奖励对当前决策的重要 性,γ越大表示智能体在采取动作a是越倾向于考虑未来奖励的影响;

[0016] 步骤4、当$Q(s,a)_t$收敛于最优映射时,认为系统已经完成了学习过程,即 Q-learning算法得到的最优策略为:

[0017]
$$a^* = \pi(s) = arg\max_a Q(s,a) \quad (4)$$

[0018] 式(4)中,arg max是对函数求参数的函数,即求得使得策略π(s)最大的Q 表映射,对于离散性的马尔可夫决策问题,这是使得系统策略最佳的必要条件,即对于环境所处的任意状态s,能量管理系统总能选择使得Q值最大的动作a^*。

[0019] 优选地,将新欧洲标准循环测试NEDC的循环测试工况表示为只有加速、减 速、等速和停车四种状态;通过电动汽车的参数表计算出车辆处于每种测试状态 时的电动机输出功率$P_{mot}(t)$。

[0020] 优选地,将驱动系统的功率$P_{mot}(t)$和电池组的温度ΔT(t)定义为步骤3中强 化学习模型的状态空间,记为$s(t) \triangleq (P_{mot}(t), \Delta T(t))$;

[0021] 将电动汽车的制冷系统的运行功率分配系数$k_{AC}(t)$定义为步骤3中强化学习 模型的动作空间；

[0022] 步骤3中强化学习模型的奖励函数定义为$r(t)=r_{avg}(t)-E(t)$，其中，$r_{avg}(t)$为0～t时刻智能体所获得的奖励的平均值，对于时间离散的系统奖励的平 均值$r_{avg}(t_n)=\frac{1}{n}[r(t_1)+r(t_2)+\cdots+r(t_n)]$，$r(t_n)$为$t_n$时刻智能体所获得的奖励。

[0023] 优选地，在所述Q-learning算法的基础上引入8-greedy策略，使得系统不 仅会采取由Q表中所得出的最优动作，并且会以某个概率对当前Q值并非最大 的动作进行试探。

[0024] 根据如上所述的本发明的方案，对电动汽车的能源管理系统而言，其需要事 先有一定行驶工况的积累，根据不同工况下汽车能源消耗的情况，环境会给予管 理系统的智能体反馈，评价先前的能源分配策略是否足够高效。经过足够多次的 训练，可以使得电动汽车能源管理系统产生一张“工况-最佳动作”的映射Q表，即得到了车辆在某一工况下的最佳能量管理策略。根据这一策略进行能量分配，将显著提高电动汽车的能量利用率。

附图说明

[0025] 图1是一个本实施例方案的电动汽车能源管理系统部件结构示意图；

[0026] 图2是一个实施例的能源管理系统能量流动示意图；

[0027] 图3是一个实施例的磷酸铁锂电池组不同温度下的放电效率示意图；

[0028] 图4是一个本实例的强化学习系统模型示意图；

[0029] 图5是一个本实例的能源管理系统Q-learning算法流程示意图。

具体实施方式

[0030] 下面结合具体实施例，进一步阐述本发明。应理解，这些实施例仅用于说明 本发明而不用于限制本发明的范围。此外应理解，在阅读了本发明讲授的内容之 后，本领域技术人员可以对本发明作各种改动或修改，这些等价形式同样落于本 申请所附权利要求书所限定的范围。

[0031] 除非另有定义，本文所使用的所有的技术和科学术语与属于本发明的技术领 域的技术人员通常理解的含义相同。本文中在本发明的说明书中所使用的术语只 是为了描述具体的实施例的目的，不是旨在于限制本发明。本文所使用的术语“或 /及”包括一个或多个相关的所列项目的任意的和所有的组合。

[0032] 如图1所示，电动汽车的主要电力设备可分为高压供电网网络、低压供电网 络、充电网络三大类。其中高压供电网络包含热管理系统和驱动系统；低压供电 网络用以支持全车低压电子部件的用电，包括灯光系统、仪表盘、ESP助力转向 系统、空调风机等；充电系统用于从电网向车载电池补充能量，其包括适用于普 通家用充电桩的交流充电系统和适用于超级充电站的直流快充系统两部分。

[0033] 据上述图1的部件结构，可得到整车能量流向示意图，如图2所示。它描述 了整车电池组与其他模块部件间的能量交换关系。其中，P_{bat}为当前电池的最大 输出功率；P_{mot}表示向驱动系统输出的最大功率；P_{AC}表示分配给空调系统的功 率，由于车厢控温系统和电池组热管理系统共用压缩机，因此可以将空调系统等 效为热管理系统；P_{LV}表示低压供电网络

所需的总功率;P_{cha}表示充电系统在当 前状态下的最大输入功率。

[0034] 在行车过程中:车辆不处于充电状态,因此$P_{cha}=0$;低压供电网络所需功 率相对于高压供电网络而言较低,且波动幅度不大,不妨将P_{LV}视为一定值;而P_{AC}和P_{mot}是能源管理系统可以根据车辆的不同工况而动态调整的量。设在任意时刻, 可被能源管理系统所分配的高压供电网络的总功率为P_{HV},即$P_{HV}(t)=P_{mot}(t)+ P_{AC}(t)$。此外,$P_{HV}$还满足:在$P_{HV}(t)=P_{bat}(t)-P_{LV}(t)-P_{cha}(t)$。显然,整车 的控制核心在于制定合理的能量分配策略,以将$P_{HV}(t)$合理地分配给$P_{mot}(t)$与 $P_{AC}(t)$,以提高车辆的能源利用率。

[0035] 在本实施例中,需上述所示的能源管理系统模型的基础上,引入一热管理系 统功率系数$k_{AC}(t)$,它表示将高压供电系统分配给热管理系统的功率占热管理系 统峰值功率的比例,取值范围为[0,1]。即分配给热管理系统的功率为:

[0036] $P_{AC}(t)=k_{AC}(t)\times P_{AC-max}$,其中$P_{AC-max}$为制冷设备(压缩机)的峰值功率。 相应地,驱动系统可获得的最大功率为:

[0037] $P_{mot}(t)=P_{HV}(t)-P_{AC}(t)=P_{HV}(t)-k_{AC}(t)P_{AC-max}$。

[0038] 能源消耗率(Energy Consumption Ratio,简称ECR)是表征汽车经济性能 的指标,工程中ECR的常用单位为kW·h/100km,它表示电动汽车在行驶100km 路程时所需要消耗的电池组的能量。此外,还应确保分配给驱动系统的功率 $P_{mot}(t)$不小于驾驶员所期望的驱动系统实际输出功率$P_{drv}(t)$,以满足车辆正常 行驶的动力需求。

[0039] 因此,在本实施例中,我们可以建立电动汽车能源管理系统的数学模型如下:

[0040] $$\begin{cases} \min\{ECR|k_{AC}(t)\}, t\in(0,T) \\ P_{mot}(t)=P_{HV}(t)-k_{AC}(t)P_{AC-max}(t) \\ P_{mot}(t)\geq P_{drv}(t), t\in(0,T) \end{cases}。$$

[0041] 在本实施例的一个具体示例中,我们对一种如今电动汽车常用的磷酸铁锂动 力电池进行探讨。磷酸铁锂动力电池具有明显的放电热效应,即电池组的温度变 化将导致电池的实际输出能量发生变化。电池放电效率是指在一定温度条件下, 电池由满电状态放电至电池截至电压的实际输出的能量与电池额定能量之比。将 电池在各温度下完全放电,可以测得在该温度下所对应的电池放电效率。e(ΔT)即 为磷酸铁锂电池组“温度-效率”函数,如图3所示。

[0042] 从图3中可以看出,电池的放电效率和电池组温度有着极大的关系。只有当 电池组温度低于44.6℃时,电池效率才高于80%。为了降低行驶周期内的总电 量消耗E,以获得较低的ECR,我们需要在提高电池的放电效率。由于NEDC循 环采用的标准测试工况下温度为20~30℃,且考虑到电池组在工作中的升温和 安全工作的温度上限,因此我们只对20~60℃温度区间的e(ΔT)函数进行多项式 拟合:

[0043] $$\begin{cases} e(\Delta T)=0.0018(\Delta T)^3-0.2573(\Delta T)^2+10.4227\Delta T-32.7208 \\ R^2=0.9964 \end{cases}, \Delta T\in[20,60]$$

[0044] 上式中,ΔT是当前的电池组温度,单位为℃;e(ΔT)是该温度下的电池放电 效率;R^2是该拟合的相关系数。

[0045] 由于电池组温度变化而耗散的能量称为电池的“无功损耗”能量,它是关于电 池组温度的函数,我们用$E_{loss}(\Delta T)$表示。已知电池的额定容量,根据式3.6我们 可推导出$E_{loss}(\Delta T)$的表示式:

[0046] $$E_{loss}(\Delta T)=\frac{e(\Delta T)-e(\Delta T_0)}{100\%}\cdot E_{bat}$$

[0047] 电动汽车在[0,t]时间段内总电量消耗E(t)可表示为：

[0048] $$E(t)=E_{mot}(t)+E_{AC}(t)+E_{LV}(t)+E_{loss}(\Delta T(t))$$
$$=\int_0^t[P_{mot}(\tau)+k_{AC}(\tau)\cdot P_{AC-max}+P_{LV}(\tau)]\,d\tau+E_{loss}(\Delta T(t))$$

[0049] 上式中，$E_{mot}(t)=\int_0^t P_{mot}(\tau)d\tau$ 是驱动系统所消耗总能量，$E_{AC}(t)=\int_0^t k_{AC}(\tau)\cdot P_{AC-max}d\tau$是热管理系统所消耗的能量，$E_{LV}(t)=\int_0^t P_{LV}(\tau)d\tau$是车辆低 压供电网络所消耗的能量，$E_{loss}(\Delta T(t))$是0～t时间段内电池组温度变化而耗散的 能量。

[0050] 综上所述，在该具体实例中求解ECR最小化的问题，可以转化为求解0～t时 间段内式上式E(t)最小化的问题。

[0051] 在本实例中，我们使用强化学习的方法对能耗最小化的问题进行求解。

[0052] 智能体在进行强化学习的过程中，设当前的环境状态s(t)，此时智能体对环 境施加动作a(t)，当环境进入到下一状态s(t+1)时，环境会向智能体反馈奖励 ——正向的奖励将对智能体执行的动作a(t)产生增强的趋势，未来再遇到这一状 态s(t)时，智能体将有更大概率采取这一动作；反之，负向奖励将降低智能体对 动作a(t)选择的偏好。图4描述了强化学习系统进行学习的基本过程。

[0053] 电动汽车的能源管理是关于时间的决策，故而可用马尔可夫决策过程的基本 方法来解决。马尔可夫决策过程包括以下几个基本元素：状态空间S，动作空间A， 状态转移概率和奖励函数r(t)。需要注意的是，由于我们的状态变量是关于时间 t的函数，因此状态转移概率为隐式。下文将更详细地阐述强化学习算法在优化 电动汽车能源消耗率问题的具体过程。

[0054] 状态空间是能源管理所有可能状态的集合。由前述分析可得出，能源管理系 统分配给驱动系统的功率$P_{mot}(t)$和电池组的温度ΔT(t)会影响能量分配策略，因 此状态空间可表示为：$s(t)\triangleq\left(P_{mot}(t),\Delta T(t)\right)$。这些表征状态的具体数值在每个 时隙开始前对于智能体而言都是未知的，智能体只能从环境中获取当前状态的信 息。

[0055] 对于一个离散的强化学习控制系统，其动作空间的取值是有限的，对于这一 实例而言即制冷系统的运行功率是一系列离散可调的量，它可以在完全关闭和全 功率开启之间选择某一中间状态进行制冷工作。为此，我们需要考虑马尔可夫决 策的每一个中间过程，即每个优化周期内的热管理系统运行功率权值，记为 $k_{AC}(t)\in A$，其中A即为动作空间。根据每一时隙内总能耗最低的需求，我们可 以确定最佳的能源分配策略。

[0056] 奖励函数是环境对智能体的反馈信号，它对智能体在该状态下执行该动作的 好坏进行评价，智能体则根据得到的奖励对自己的策略进行调整。

[0057] 在时刻t，移动设备在状态$s\in S$采取行动$a\in A$后，将获得奖励，使得移动设 备知道它采取的动作是否正确，环境的正向奖励将对能源管理系统采取的某一动 作a具有加强作用，未来在同样的环境条件下，采用这个动作的偏好会有所增加； 反之，如果得到负向奖励，那么以后产生这个动作的趋势就会减弱。能源管理系 统根据获得的奖励以学习的方式

不断更新从状态到动作的映射策略，最终达到能 耗最小化的目的。

[0058] 为避免保守的过程导致的智能体学习效率降低，我们需要对即使奖励$r_1(t)$进 行改进。设$r_{avg}(t)$为0～t时刻智能体所获得的奖励的平均值，则$r_{avg}(t)$可以作为 一个参考值用于比较当前获得的即时奖励高于或低于以往获得奖励的平均值，从 而获得对当前执行动作的评价。具体而言，我们将实际进行运算的奖励定义为：

[0059] $r(t)=r_{avg}(t)-E(t)$

[0060] 对于时间离散的系统，奖励的平均值可被表示为：

[0061] $$r_{avg}(t_n)=\frac{1}{n}[r(t_1)+r(t_2)+\cdots+r(t_n)]$$

[0062] 前述部分，我们已经交代了实例中强化学习的状态空间、动作空间和奖励函 数，接下来将使用Q-learning算法(一种强化学习算法)对能量分配策略进行学 习和优化。该策略使得车辆的能源管理系统拥有一个状态动作值矩阵，即Q表。Q表内的每一项Q(s,a)为状态s与动作a的映射关系。当电动汽车能源管理系统在 某一状态s下进行能量分配时，矩阵Q将采用下式的方法进行更新：

[0063] $$Q(s,a)_{t+1}\leftarrow Q(s,a)_t+\alpha\left[r_t+\gamma\max_{a(t+1)}Q(s,a)_{t+1}-Q(s,a)_t\right]$$

[0064] 其中，参数$\alpha\in(0,1]$是学习率，它表示Q值的过去值和新获得奖励的加权关 系；$\gamma\in[0,1]$衰减系数反应未来奖励对当前决策的重要性，该值越大表示智能体 在采取动作a是越倾向于考虑未来奖励的影响。采取行动a(t)后，环境从s(t)变 为s(t+1)，此时能源管理系统对动作值矩阵Q进行更新。其中，$r_t+\gamma\max_{a(t+1)}Q(s,a)_{t+1}$部分是Q(s,a)的现实值，而$Q(s,a)_t$是在采取动作a之前， 智能体对映射(s,a)的估计。对现实值和估计值求差再乘以学习效率α，经过若干 次学习，则Q表的中的估计值与现实值的差距将逐渐减小并最终收敛获得一组 稳定的关于(s,a)的映射。当Q(s,a)收敛于最优映射时，我们可以认为系统已经完成了学习过程，即Q-learning算法得到的最优策略为：

[0065] $$a^*=\pi(s)=arg\max_a Q(s,a)$$

[0066] 上式中，arg max是对函数求参数的函数，即求得使得策略π(s)最大的Q表 映射。对于离散性的马尔可夫决策问题，这是使得系统策略最佳的必要条件，即 对于环境所处的任意状态s，能量管理系统总能选择使得Q值最大的动作a^*。图 5展现了能源管理系统Q-learning算法的执行过程。

[0067] 在本实例Q-learning算法的基础上，我们还需要引入ε-greedy策略，来防 止次优解的Q值积累，而导致系统无法选择到最优解的情况发生。该策略不仅 会采取Q值最大的“最优”a(t)，并且会以某个概率对当前Q值并非最大的动作进 行试探。具体而言，ε-greedy策略将以$\varepsilon\in[0,1]$的概率选择当前Q值最大的动作， 而有1-ε的概率将在Q表中随机选取动作。该策略保证了经过足够多次试错试 验后，系统总能使得当前状态的最优动作的Q值最大。

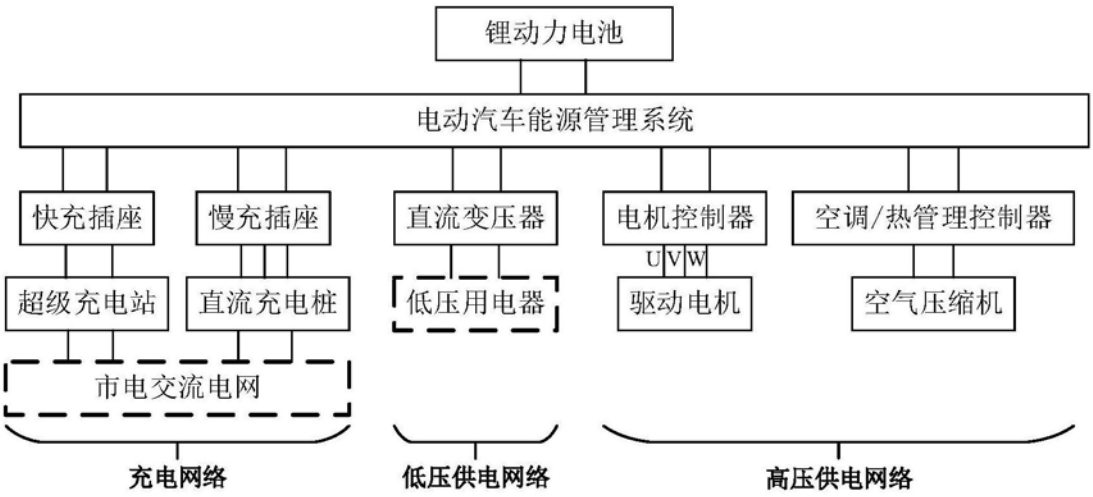

锂动力电池
电动汽车能源管理系统
快充插座
慢充插座
直流变压器
电机控制器
空调/热管理控制器
U V W
超级充电站
直流充电桩
低压用电器
驱动电机
空气压缩机
市电交流电网
充电网络
低压供电网络
高压供电网络

图1

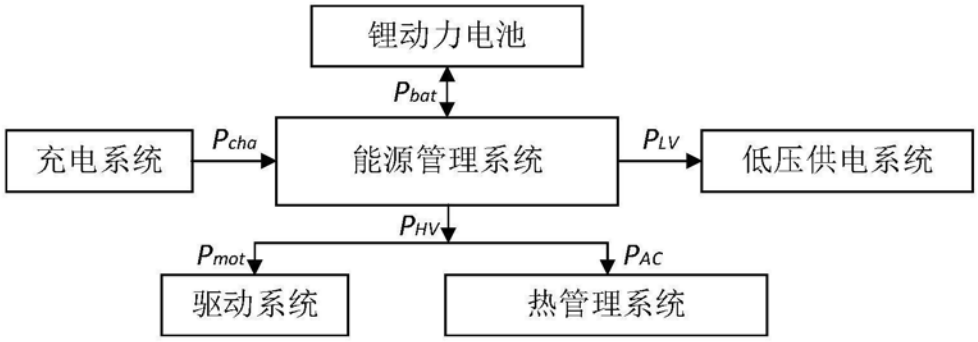

锂动力电池
P_{bat}
充电系统
P_{cha}
能源管理系统
P_{LV}
低压供电系统
P_{HV}
P_{mot}
P_{AC}
驱动系统
热管理系统

图2

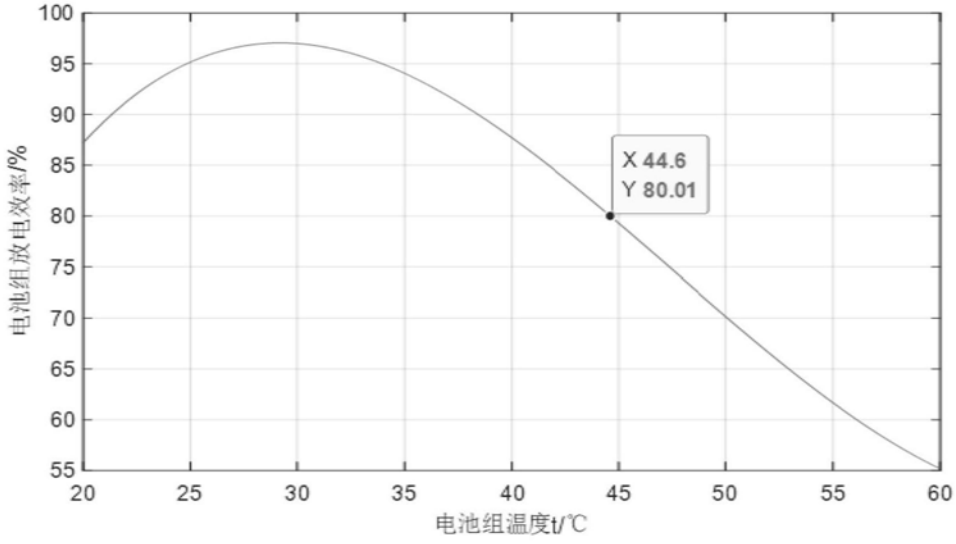

电池组放电效率/%
电池组温度t/℃
X 44.6
Y 80.01
100
95
90
85
80
75
70
65
60
55
20
25
30
35
40
45
50
55
60

图3

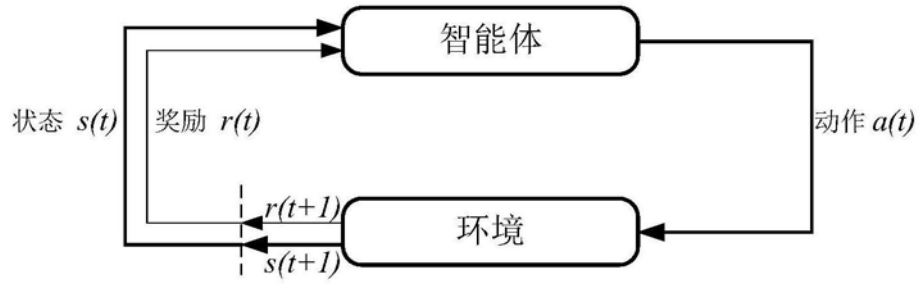


图4

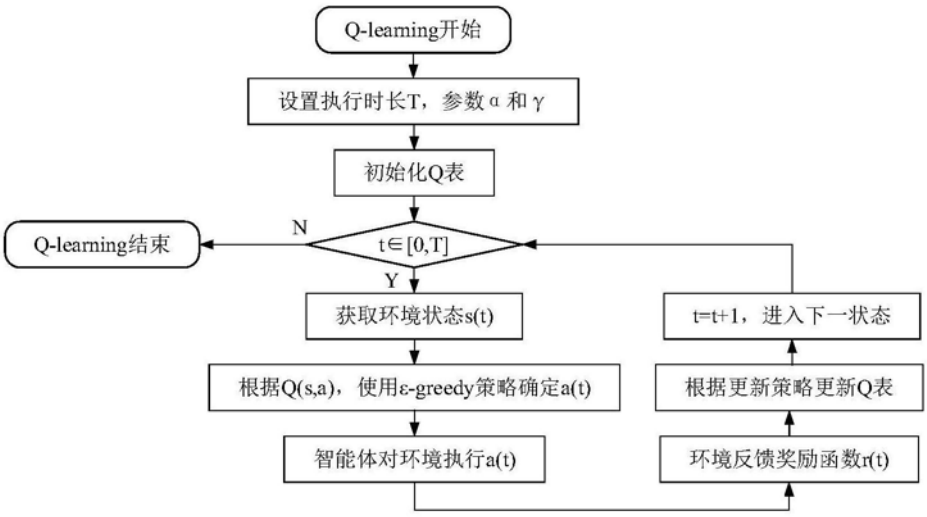


图5