



(12) 发明专利

(10) 授权公告号 CN 103217899 B

(45) 授权公告日 2016. 05. 18

(21) 申请号 201310036976. X

(22) 申请日 2013. 01. 30

(73) 专利权人 中国科学院自动化研究所

地址 100190 北京市海淀区中关村东路 95
号

(72) 发明人 赵冬斌 朱圆恒 刘德荣

(74) 专利代理机构 中科专利商标代理有限责任
公司 11021

代理人 宋焰琴

(51) Int. Cl.

G05B 13/04(2006. 01)

(56) 对比文件

US 7047224 B1, 2006. 05. 16,

US 6532454 B1, 2003. 05. 11,

CN 101789178 A, 2010. 07. 28,

Huaguang Zhang, Yanhong Luo, Derong

Liu. Neural-network-based near-optimal

control for a class of discrete-time affine nonlinear systems with control constraints. 《IEEE TRANSACTIONS ON NEURAL NETWORKS》. 2009, 第 20 卷 (第 9 期),

赵冬斌, 刘德荣, 易建强. 基于自适应动态规划的城市交通信号优化控制方法综述. 《自动化学报》. 2003, 第 35 卷 (第 6 期), 676-681.

审查员 李亚琼

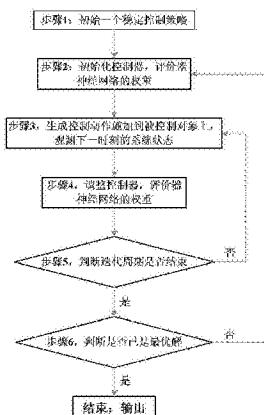
权利要求书2页 说明书5页 附图3页

(54) 发明名称

基于数据的 Q 函数自适应动态规划方法

(57) 摘要

本发明提出一种基于数据的 Q 函数自适应动态规划方法解决最优控制问题。所述方法主要包括: 步骤 1, 初始化稳定的控制策略; 步骤 2, 用已有的控制策略初始化控制器 (actor)、评价器 (critic) 神经网络的权重; 步骤 3, 根据当前控制策略和当前时刻系统状态, 生成被控制系统的控制动作并施加到被控制对象上, 观测下一时刻的系统状态; 步骤 4, 调整控制器、评价器神经网络的权重; 步骤 5, 判断当前迭代周期是否已经结束, 是则进入步骤 6, 否则回到步骤 3; 步骤 6, 判断最近两个迭代周期产生的神经网络权重是否有明显变化, 是则用新产生的控制器、评价器神经网络进入步骤 2, 否则输出最终的控制器神经网络控制器。



1. 一种通过自适应动态规划优化系统控制策略的方法，其包括以下步骤：

步骤1，初始化任意一个稳定的控制策略作为当前控制策略；

步骤2，使用当前控制策略初始化控制器、评价器神经网络的权重；

步骤3，根据当前控制策略和当前时刻被控系统的状态，生成控制动作并施加到被控系统上，获得下一时刻的系统状态；

步骤4，根据前一时刻系统状态、相应控制动作和下一时刻的系统状态，调整控制器、评价器神经网络的权重，获得调整后的控制器和评价器神经网络权重；

步骤5，判断当前迭代周期是否已经结束，是则进入步骤6，否则将调整后的控制器神经网络权重对应的控制策略作为当前控制策略返回步骤3继续执行；

步骤6，判断最近两个迭代周期所产生的控制器、评价器神经网络权重是否有明显变化，是则将调整后的控制器神经网络对应的控制策略作为当前控制策略进入步骤2继续优化，否则输出当前控制器神经网络对应的控制策略作为最优的控制策略；

其中，步骤4中调整评价器神经网络的权重的公式表示如下：

$$z(j) = r(x_k, u_k)$$

$$h(j) = \Phi(x_k, u_k) - \Phi(x_{k+1}(u_k), u_{k+1}^{(i)})$$

$$l(j) = P(j-1)h(j)[h(j)^T P(j-1)h(j) + 1]^{-1}$$

$$P(j) = [I - l(j)h(j)^T]P(j-1)$$

$$W_Q^{(i,j)} = W_Q^{(i,j-1)} + I(j)[z(j) - h(j)^T W_Q^{(i,j-1)}]$$

其中， $z(j)$ 、 $h(j)$ 、 $l(j)$ 和 $P(j)$ 为中间变量，效用函数 $r(\cdot, \cdot)$ 定义为 $x_k^T Q x_k + u_k^T R u_k$ ； Q 和 R 是正定矩阵； $x_{k+1}(u_k)$ 是指在系统状态 x_k 下施加控制动作 u_k 后系统下一时刻的状态； $\Phi(x_k, u_k)$ 和 $\Phi(x_{k+1}(u_k), u_{k+1}^{(i)})$ 是激活函数， $u_{k+1}^{(i)}$ 指当前控制策略下在系统状态为 $x_{k+1}(u_k)$ 时对应的控制动作； $W_Q^{(i,j)} = [w_Q^{(i,j)}_1, w_Q^{(i,j)}_2, \dots, w_Q^{(i,j)}_N]^T$ 是第*i*次迭代周期内进行的第*j*次调整后的评价器神经网络的权重， I 是单位阵；第*j*次调整对应第*k*时刻，第*j-1*次调整对应第*k-1*时刻； M 和 N 分别是控制器神经网络和评价器神经网络的隐含层神经元个数。

2. 根据权利要求1所述的方法，其特征在于，该方法基于被控系统的系统状态和控制动作的性能指标函数，通过在线运行被控系统实时优化所述被控系统的控制策略。

3. 如权利要求1所述的方法，其特征在于，步骤4中控制器神经网络的权重如下调整：

$$W_u^{(i+1,j+1)} = W_u^{(i+1,j)} - \alpha \frac{\Psi}{(\Psi^T \Psi + 1)^{0.5}} \cdot \frac{\nabla_u \Phi^T}{(\nabla_u \Phi^T \nabla_u \Phi + 1)^{0.5}} \cdot W_Q^{(i,j)}$$

其中， α 被称为下降因子； $W_u^{(i+1,j+1)}$ 表示第*i+1*轮迭代周期第*j+1*次调整后的评价器神经网络的权重； $\Psi = \Psi(x_k)$ 为激活函数， $\nabla_u \Phi = \left. \frac{\partial \Phi}{\partial u_k} \right|_{(x_k, \bar{u}_k)}$ ， \bar{u}_k 是被施加到被控系统上的控制动作， $W_Q^{(i,j)}$ 表示第*i*次迭代周期内进行了第*j*次调整后的评价器神经网络的权重，第*j*次调

整对应第k时刻,第j-1次调整对应第k-1时刻。

4. 如权利要求1所述的方法,其特征在于,所述控制策略和性能指标函数如下表示:

$$\hat{u}^{(i)}(x_k) = W_u^{(i)T} \Psi(x_k)$$

$$\hat{Q}^{(i)}(x_k, u_k) = W_Q^{(i)T} \Phi(x_k, u_k)$$

其中,上标i指的是第i次迭代周期; \hat{Q} 和 \hat{u} 分别表示性能指标函数和控制策略;
 $W_u^{(i)} = [w_{u,1}^{(i)}, w_{u,2}^{(i)}, \dots, w_{u,M}^{(i)}]^T$ 和 $W_Q^{(i)} = [w_{Q,1}^{(i)}, w_{Q,2}^{(i)}, \dots, w_{Q,N}^{(i)}]^T$ 分别是控制器和评价器
神经网络的权重, $\Psi = \Psi(x_k)$ 和 $\Phi(x_k, u_k)$ 为激活函数。

5. 如权利要求1所述的方法,其特征在于,通过所述控制器神经网 络计算得到当前时
刻要施加到被控系统上的控制动作,具体如下表示:

$$\bar{u}_k = \hat{u}^{(i)}(x_k) + n_k$$

其中, n_k 是探索噪声, $\hat{u}^{(i)}(x_k)$ 是当前控制策略下在系统状态 x_k 时计算得到的控制动作。

6. 如权利要求1所述的方法,其特征在于,步骤4中通过建立的Q函数迭代更新调整所述
控制器、评价器神经网络的权重,所述Q函数如下表示:

$$\hat{Q}^{(i)}(x_k, u_k) = r(x_k, u_k) + \hat{Q}^{(i)}(x_{k+1}(u_k), u_{k+1}^{(i)}), \hat{Q}^{(i)}(0, 0) = 0$$

$$\hat{u}^{(i+1)} = \arg \min_{\mu_k} \hat{Q}^{(i)}(x_k, \mu_k)$$

其中, $\hat{Q}^{(i)}$ 是第i次迭代周期的性能指标函数, $\hat{u}^{(i+1)}$ 是第i+1次迭代周期的控制策略,效
用函数 $r(\cdot, \cdot)$ 定义为 $x_k^T Q x_k + u_k^T R u_k$, Q和R是正定矩阵, $x_{k+1}(u_k)$ 指在采用 u_k 这个控制动
作后的系统状态, μ_k 指在系统状态 x_k 时可以采用的任意一种控制动作。

7. 如权利要求1所述的方法,其特征在于,步骤5中通过判断是否达到参数调整的最大
次数来判断当前迭代周期是否已经结束。

基于数据的Q函数自适应动态规划方法

技术领域

[0001] 本发明涉及智能控制技术领域,尤其涉及基于数据的Q函数自适应动态规划方法。

背景技术

[0002] 在工业生产、航空航天、汽车工程等领域,被控对象能够在有限的资源下使用最小的资源来完成控制目标,即最优控制。最优控制指的是找到一个最优控制策略能够使得性能指标函数达到最优。性能指标函数是与系统的状态和所采用的控制策略有关,它能够反映该控制策略在当前以及以后时刻的控制效果。针对离散系统性能指标函数可以用数学形式可以表示成如下式子:

$$[0003] V(x_k) = \sum_{n=k}^{\infty} r(x_n, u_n)$$

[0004] 其中系统运行时间用下标k表示, x_k 和 u_k 分别指k时刻系统状态和控制动作。 $r(\cdot, \cdot)$ 被称为效用函数,反映某一时刻当前系统的运行好坏。因此最优控制便是寻找使上式有最优解的控制策略,即:

$$[0005] V^*(x_k) = \min_{\mu} \sum_{n=k}^{\infty} r(x_n, \mu_n)$$

[0006] μ 指的是任意控制策略。最优控制是现代控制理论中重要的组成部分。然而,由于计算的复杂性,最优的控制策略一般是无法直接计算得到的。尤其是针对非线性系统,计算难度非常巨大。特别是车辆行驶的车道保持问题,不仅要考虑控制车辆在保持在车道内,还要使控制动作尽可能小,控制时间尽可能短,是典型的非线性系统的最优控制问题。而且,在实际应用中,由于车内乘坐人员重量变化、路况变化等,很难得到精确的车辆模型,提出了基于数据的最优控制器的设计问题。

[0007] 自适应动态规划自20世纪80年代提出来,得到了快速的发展。它主要是用来解决动态规划问题,尤其是在求解最优控制方面表现了巨大的优势。自适应动态规划方法一般使用控制器-评价器(actor-critic)结构和神经网络,用来逼近性能指标函数和控制策略,采用迭代的方法逐步逼近,最终收敛到最优性能指标函数和最优控制策略。

[0008] 然而,传统的自适应动态规划方法一般是逼近仅和系统状态有关的V函数。V函数相对比较简单,计算方便,但是V函数自适应动态规划方法的运行依赖系统模型因而常常被用于离线运行。当系统模型未知时,V函数自适应动态规划方法将不再适用,除非再加上一个系统辨识网络用来辨识系统模型。但加上系统辨识网络后,整个算法的结构变得复杂、冗余,而且辨识网络的训练和V函数自适应动态规划方法的运行是完全分开的,这不利于整个算法。因此提出一种不依赖于系统模型的自适应动态规划方法显得尤为重要。

发明内容

[0009] 针对传统的自适应动态规划依赖系统模型,该发明提出一种基于Q函数的自适应动态规划方法,用于解决一类非线性系统的最优控制问题,并给出了车辆行驶中的车道保

持问题的具体实施方式。定义的Q函数不仅与系统状态有关,同时也与控制动作相关,使得Q函数能够包含系统模型信息,因而Q函数自适应动态规划方法不依赖系统模型,而是基于实时产生的系统状态和相应的控制动作来调整控制器和评价器神经网络的权重。最终,Q函数自适应动态规划方法能够在线运行并使得控制器和评价器神经网络最终迭代收敛到最优控制策略和最优性能指标函数。特别适用于线性或非线性离散系统的在线求解最优控制问题。该方法可以成功地应用在车道保持问题上。

[0010] 本发明提出一种通过自适应动态规划优化系统控制策略的方法,其包括以下步骤:

- [0011] 步骤1,初始化任意一个稳定的控制策略作为当前控制策略;
- [0012] 步骤2,使用当前控制策略初始化控制器、评价器神经网络的权重;
- [0013] 步骤3,根据当前控制策略和当前时刻被控系统的状态,生成控制动作并施加到被控系统上,获得下一时刻的系统状态;
- [0014] 步骤4,根据前一时刻系统状态、相应控制动作和下一时刻的系统状态,调整控制器、评价器神经网络的权重,获得调整后的控制器和评价器神经网络权重;
- [0015] 步骤5,判断当前迭代周期是否已经结束,是则进入步骤6,否则将调整后的控制器神经网络权重对应的控制策略作为当前控制策略返回步骤3继续执行;
- [0016] 步骤6,判断最近两个迭代周期所产生的控制器、评价器神经网络权重是否有明显变化,是则将调整后的控制器神经网络对应的控制策略作为当前控制策略进入步骤2继续优化,否则输出当前控制器神经网络对应的控制策略作为最优的控制策略。
- [0017] 本发明直接利用实时采集的数据,不依赖于系统模型。将车道保持作为本发明的研究对象,如图2所示。控制目标是控制前轮转角使得车辆能够稳定运行在车道中央。
- [0018] 综上所述,与传统的自适应动态规划方法相比,本发明提出的Q函数自适应动态规划方法具有以下优点:
- [0019] ●本发明提出的Q函数自适应动态规划方法不依赖于被控对象模型,而是基于采集的系统数据,使得该方法适用于在线运行;
- [0020] ●不论是线性还是非线性离散系统,该方法都能够适用;
- [0021] ●采用策略迭代的方法,保证整个算法在运行中,控制策略始终都是稳定的且能收敛到最优解。
- [0022] ●控制动作加入了探索噪声,既满足了持续激励条件,同时也保证了整个系统在运行当中不断输出有用的系统数据。

附图说明

- [0023] 图1是本发明中基于数据的Q函数自适应动态规划方法流程图;
- [0024] 图2是本发明优选实施例中车道保持问题示意图;
- [0025] 图3是本发明中控制器-评价器结构图;
- [0026] 图4是本发明中控制器和评价器的神经网络结构示意图。

具体实施方式

- [0027] 为使本发明的目的、技术方案和优点更加清楚明白,参照附图,对本发明进行进一

步详细说明。

[0028] 图1是基于Q函数自适应动态规划方法的应用流程图。

[0029] 如图1所示,该方法包括以下几个步骤:

[0030] 步骤1,首先初始化任意一个稳定的控制策略,要求这个控制策略能够稳定控制被控系统。

[0031] 图2是车道保持问题示意图。其中车辆重心横向偏移距离 y_{cg} 指的是车辆重心到车道的偏移距离,车辆与车道的偏转角 ψ_d 指的是车辆方向与车道切线方向的夹角,而 δ 则是前轮转角。稳定的控制策略指的是在某一区域内,在任意初始状态下,控制策略能够对被控系统进行稳定控制。初始稳定的控制策略不仅保证了相应的性能指标函数是有效的,同时有利于Q函数自适应动态规划方法的在线运行。初始的稳定控制策略不需要是最优的,可以是任意一种稳定的控制策略。在实际应用中,一个被控系统的稳定控制策略是很容易得到,如常见的LQR方法、模糊控制等等都可以作为初始的稳定控制策略。在车道保持问题上,稳定的控制策略即是能够将车辆稳定行驶在车道上的控制策略。

[0032] 步骤2,采用控制器-评价器结构,并用神经网络逼近控制策略和性能指标函数。用已有的控制策略初始化控制器、评价器神经网络的权重进入一个迭代周期。

[0033] 图3是控制器-评价器结构图,示出了评价器、控制器和被控系统之间的数据流向,其中 u_k 和 x_k 分别表示控制动作和该控制动作下的系统状态。图4是神经网络结构图。神经网络结构包括输入、n个隐藏神经元和相应的n个神经元权重 w_1, w_2, \dots, w_n 和输出。控制器和评价器神经网络分别用来逼近控制策略和性能指标函数。控制器神经网络用来计算控制动作,而评价器神经网络则用来反映当前控制策略的性能指标,从而改进当前控制策略。神经网络逼近控制策略和性能指标函数可以用如下公式表示:

$$\hat{u}^{(i)}(x_k) = W_u^{(i)T} \Psi(x_k) \quad (1)$$

$$\hat{Q}^{(i)}(x_k, u_k) = W_Q^{(i)T} \Phi(x_k, u_k) \quad (2)$$

[0036] 其中,上标i指的是第i次迭代周期; \hat{Q} 和 \hat{u} 分别表示由神经网络逼近的性能指标函数和控制策略。 $W_u^{(i)} = [w_{u,1}^{(i)}, w_{u,2}^{(i)}, \dots, w_{u,M}^{(i)}]^T$ 和 $W_Q^{(i)} = [w_{Q,1}^{(i)}, w_{Q,2}^{(i)}, \dots, w_{Q,N}^{(i)}]^T$ 分别是控制器和评价器神经网络的权重, $\Psi(x_k)$ 和 $\Phi(x_k, u_k)$ 为激活函数被称为激活函数,其可以为高斯函数或二次函数,而M和N则是两个神经网络的隐含层神经元个数。符号T表示对向量或矩阵作转置。输入变量包括车辆重心横向偏移距离 y_{cg} ,车辆与车道的偏转角 ψ_d ,以及车辆自身的旋转角速度 r_d 。控制动作是前轮转角 δ 。根据神经网络的逼近性,通过选取合适的神经网络结构并调整相应的神经网络权重,是可以有效的逼近性能指标函数和控制策略。尤其是当被控系统是非线性系统时,性能指标函数和控制策略是高度非线性函数,无法直接用函数表示时,神经网络能够有效地解决相应的问题。

[0037] 步骤3,根据当前控制策略和当前时刻系统状态,生成控制动作并施加到系统上,观测下一时刻的系统状态。

[0038] 为了满足持续激励条件,用上面所述的控制器神经网络计算得到的控制动作需要加上一个探索噪声才可以施加到被控系统上:

[0039] $\bar{u}_k = \hat{u}^{(i)}(x_k) + n_k$ (3)

[0040] 其中 n_k 指的是探索噪声, \bar{u}_k 是最终被施加到被控系统上的控制动作, 在车道保持时线型变换为前轮转角 δ 。

[0041] 步骤4, 根据已有的系统观测量, 包括前一时刻系统状态、相应控制动作和下一时刻的系统状态, 调整控制器、评价器神经网络的权重。由于控制策略和性能指标函数是随着控制器、评价器神经网络的权重而改变的, 调整了控制器、评价器神经网络的权重, 意味着控制策略和性能指标函数的更新。

[0042] 为了保证该发明的有效运行, 采用策略迭代的方法, 计算当前迭代周期的控制策略的性能指标函数 \hat{Q} 和下一迭代周期的控制策略 \hat{u} :

[0043] $\hat{Q}^{(i)}(x_k, u_k) = r(x_k, u_k) + \hat{Q}^{(i)}(x_{k+1}(u_k), u_{k+1}^{(i)}), \hat{Q}^{(i)}(0, 0) = 0$ (4)

[0044] $\hat{u}^{(i+1)} = \arg \min_{\mu_k} \hat{Q}^{(i)}(x_k, \mu_k)$ (5)

[0045] 其中效用函数 $r(\cdot, \cdot)$ 定义为 $x_k^T Q x_k + u_k^T R u_k$, Q 和 R 是正定矩阵, $x_{k+1}(u_k)$ 指在采用 u_k 这个控制动作后的系统状态, $u_{k+1}^{(i)}$ 指当前控制策略下在系统状态为 $x_{k+1}(u_k)$ 时对应的控制动作, μ_k 指在系统状态 x_k 时可以采用的任意一种控制动作。这样效用函数与系统状态和控制动作相关, 从而控制目标便是找到最优的控制策略使得系统稳定时间尽可能短, 施加的控制动作尽量的小。

[0046] 由于性能指标函数 $\hat{Q}^{(i)}$ 是关于评价器神经网络权重 $W_Q^{(i)}$ 的线性函数(见公式(2)), 利用实时采集的系统观测量来调整评价器神经网络权重时, 可以采用递推最小二乘法来计算评价器神经网络权重 $W_Q^{(i)}$, 即根据公式(1)、(2)、(4)和(5)得到下面的公式表示:

[0047] $z(j) = r(x_k, u_k)$

[0048] $h(j) = \Phi(x_k, u_k) - \Phi(x_{k+1}(u_k), u_{k+1}^{(i)})$

[0049] $l(j) = P(j-1)h(j)[h(j)^T P(j-1)h(j) + 1]^{-1}$ (6)

[0050] $P(j) = [I - l(j)h(j)^T]P(j-1)$

[0051] $W_Q^{(i,j)} = W_Q^{(i,j-1)} + l(j)[z(j) - h(j)^T W_Q^{(i,j-1)}]$

[0052] 其中, 上标 j 是指在这个第 i 次迭代周期内进行第 j 次调整, j 与当前迭代周期中的时刻有关, 即当第 j 次调整对应第 k 时刻时, 第 $j+1$ 次调整对应第 $k+1$ 时刻, $z(j)$ 、 $h(j)$ 、 $l(j)$ 和 $P(j)$ 是在运行递推最小二乘法时需要的一些中间变量, $u_{k+1}^{(i)}$ 指当前控制策略下在系统状态为 $x_{k+1}(u_k)$ 时对应的控制动作。在每轮迭代周期中, 用当前时刻的权重调整下一时刻的权重, 最终得到收敛后的评价器神经网络的权重。

[0053] 在求解控制策略时, 由公式(5)无法给出一个明确的表达式来作为控制器神经网络的权重, 因此, 选择梯度下降法来计算控制器神经网络的权重 $W_u^{(i+1)}$:

$$[0054] \quad W_u^{(i+1,j+1)} = W_u^{(i+1,j)} - \alpha \frac{\Psi}{(\Psi^T \Psi + 1)^{0.5}} \cdot \frac{\nabla_u \Phi^T}{(\nabla_u \Phi^T \nabla_u \Phi + 1)^{0.5}} \cdot W_Q^{(i,j)}$$

[0055] 其中, α 被称为下降因子; $\Psi = \Psi(x_k)$ 和 $\nabla_u \Phi = \frac{\partial \Phi}{\partial u_k} \Big|_{(x_k, \tilde{u}_k)}$ 。 $(\nabla_u \Phi^T \nabla_u \Phi + 1)^{0.5}$ 和 $(\nabla_u \Phi^T \nabla_u \Phi + 1)^{0.5}$ 是用来进行归一化, 保证算法的有效运行。

[0056] 步骤5, 判断当前迭代的周期是否已经结束, 即达到最大的参数调整次数; 是则意味着生成了新的控制策略和性能指标函数, 进入步骤6, 否则回到步骤3继续调整控制器、评价器神经网络的权重。

[0057] 步骤6, 判断最近两个迭代周期产生的神经网络权重是否有明显变化, 是则表示还未得到最优解, 用新产生的控制器、评价器神经网络进入步骤2, 否则输出最终的最优控制器神经网络控制器, 如实现车道保持的最优控制器。

[0058] 经过上述步骤1~6后, 最终获得的控制器和评价器神经网络被认为是最优控制策略和最优性能指标函数。

[0059] 以上所述的方法步骤, 对本发明的目的、技术方案和有益效果进行了进一步详细说明, 凡在本发明的精神和原则之内, 所做的任何修改、等同替换、改进等, 均应包含在本发明的保护范围之内。

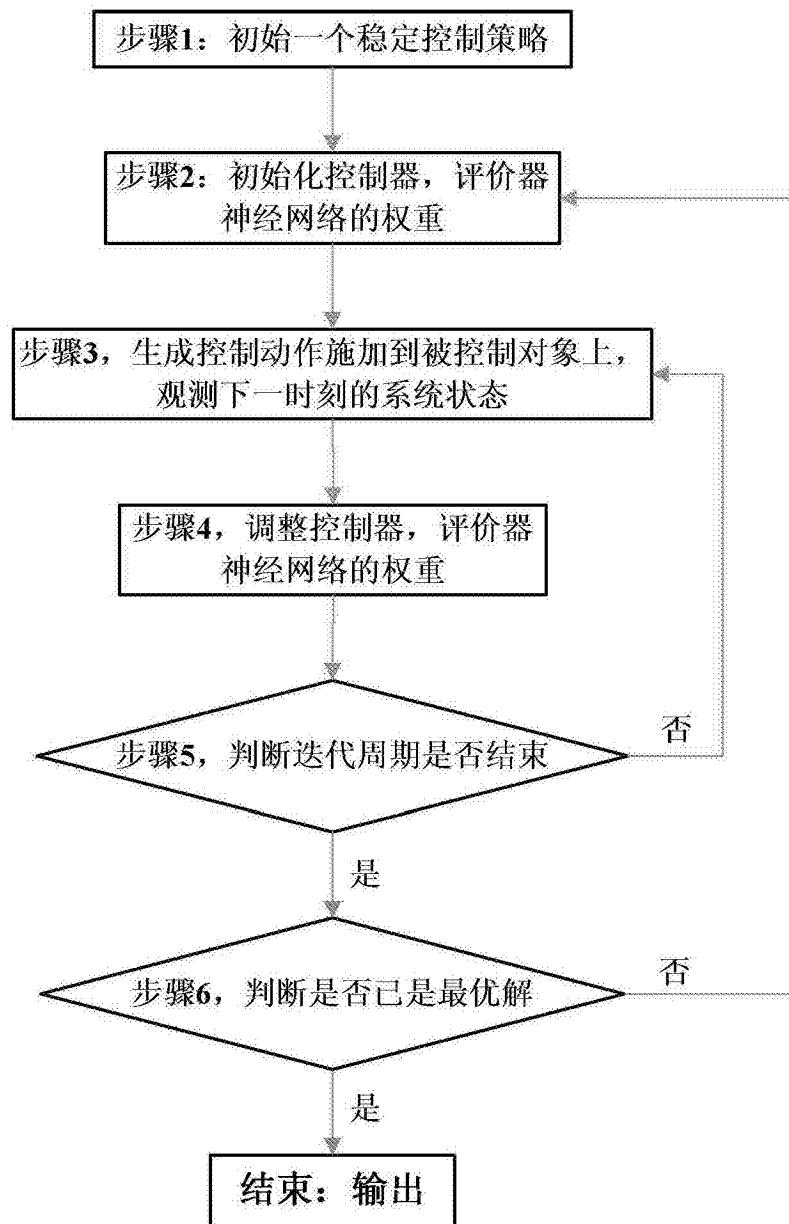


图1

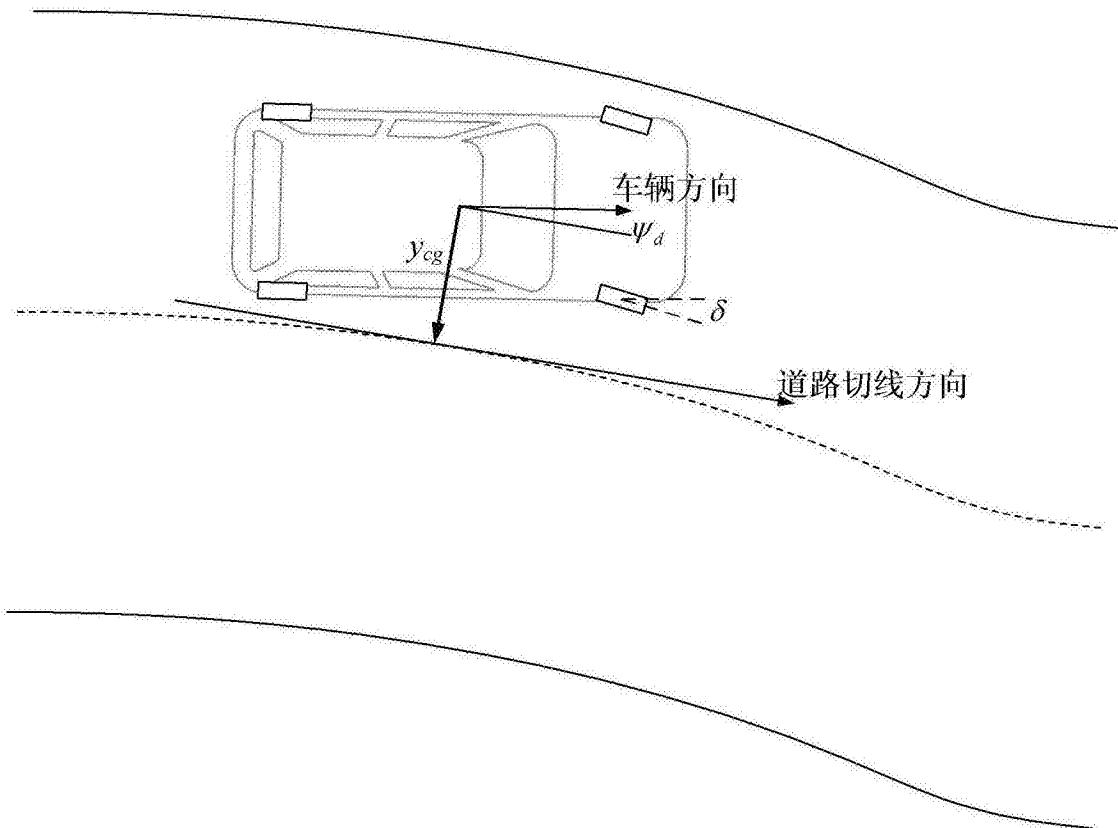


图2

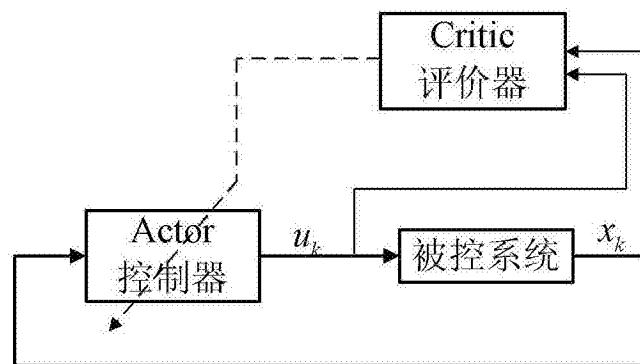


图3

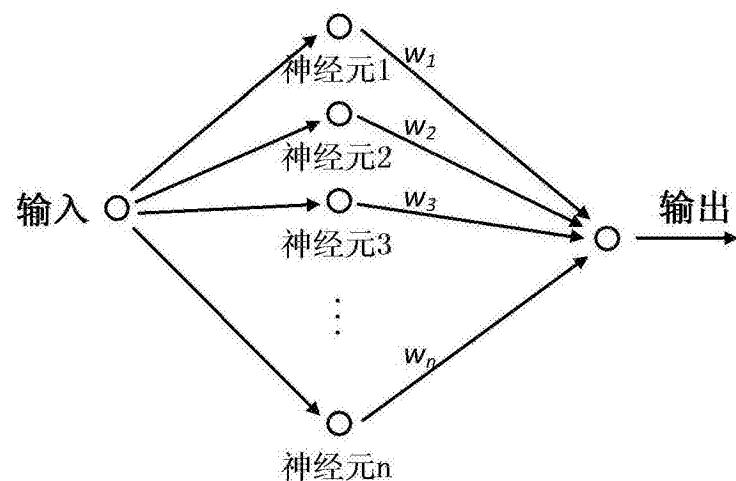


图4