(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 106097733 A
(43)申请公布日 2016.11.09

(21)申请号 201610696748.9

(22)申请日 2016.08.22

(71)申请人 青岛大学
地址 266071 山东省青岛市市南区宁夏路308号

(72)发明人 王冬青 张震 董心壮 丁军航 宋婷婷

(51)Int.Cl.
G08G 1/08(2006.01)

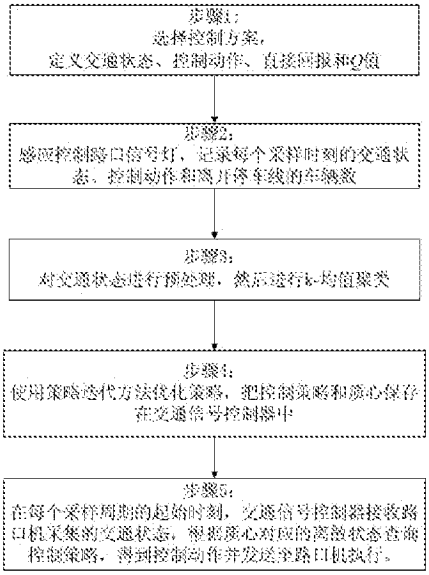
权利要求书1页 说明书7页 附图2页

(54)发明名称

一种基于策略迭代和聚类的交通信号优化控制方法

(57)摘要

本发明提出一种基于策略迭代和聚类的交通信号优化控制方法，该方法涉及智能优化技术领域，包括：步骤1，选择控制方案，定义交通状态、控制动作、直接回报和*Q*值；步骤2，感应控制交通信号灯，记录每个采样时刻的交通状态、控制动作和离开停车线的车辆数；步骤3，对交通状态进行预处理，然后进行k均值聚类；步骤4，在路口机中使用策略迭代方法优化策略，把优化得到的策略和步骤3中得到的质心保存在交通信号控制器中；步骤5，使用步骤4获得的控制策略替代感应控制，在每个采样周期的起始时刻，交通信号控制器接收路口机采集的交通状态，根据质心对应的离散状态查询控制策略，得到控制动作并发送至路口机执行。



CN 106097733 A

1.一种基于策略迭代和聚类的交通信号优化控制方法，包括以下步骤：

步骤1，选择待优化的信号控制方案为固定相序控制，定义交通状态为当前相位和下一相位的车辆排队长度，定义控制动作为保持当前相位或切换到下一相位，定义直接回报为一个与单个采样周期内离开停车线的车辆数有关的变量，定义状态-动作对为离散交通状态和控制动作组成的数据向量，定义每个状态-动作对的Q值表示处于相应离散交通状态下采取控制动作后获得的期望累积回报，定义控制策略为每个离散交通状态应该执行的控制动作；

步骤2，路口机把交通信号控制器的控制策略设置为感应控制，最小绿灯时间、最大绿灯时间设置为采样周期的正整数倍，单位绿灯延长时间与采样周期相同，路口机对交通状态、执行的相位动作和离开停车线的车辆数进行采样并记录样本，采样方法为：在每个采样时刻记录交通状态、控制动作和每个采样周期离开停车线的车辆数；

步骤3，路口机采集到指定数目的样本后，对样本中的交通状态进行离散化，离散化方法为：先对采样得到的交通状态进行归一化，并且去掉间距超过预设阈值的交通状态，再进行k-均值聚类，将得到的质心进行编号，每个质心对应一个离散交通状态，并且把归一化样本中的交通状态用最近的质心的编号表示，得到对应的离散交通状态；

步骤4，路口机使用策略迭代优化策略，把优化得到的策略和步骤3中得到的质心保存在交通信号控制器中；

步骤5，路口机设置交通信号控制器的控制策略为步骤4获得的控制策略，并把决策周期设置为采样周期，在每个决策时刻，交通信号控制器接收路口机检测到的交通状态，进行归一化，计算归一化后的交通状态到每个质心的距离，求出距离最近的质心，根据质心对应的离散交通状态查询控制策略，得到控制动作并发送至路口机执行。

2.如权利要求1所述的方法，其中，在归一化之后，k-均值聚类之前，要去掉间距超过预设阈值的交通状态，方法为：先从归一化样本中随机选择一个交通状态加入一个空的数据集，然后把归一化样本中剩余的交通状态按照下列原则加入到数据集中：如果归一化样本中的交通状态到数据集中所有交通状态的欧氏距离都大于预设阈值，则把该交通状态加入数据集，否则不加入。

一种基于策略迭代和聚类的交通信号优化控制方法

技术领域

[0001] 本发明涉及智能优化技术领域。

背景技术

[0002] 交通信号的优化控制是城市交通管理与控制系统的重要组成部分，交通信号控制策略的优劣直接影响整个路网的运输效率和人们的出行体验，因此，各种智能优化控制方法被提出并被尝试应用于交通信号控制策略的优化。

[0003] 动态规划是一种求解最优控制策略的方法，包括值迭代和策略迭代两种方法。对策略交通状态、相位和直接回报进行采样，然后利用样本对控制策略进一步优化，因而很适合解决交通信号优化控制问题。在对交通信号控制问题进行策略迭代时，需要将车辆排队长度等连续变量进行离散化。传统的离散化方法是将整个状态空间进行均一划分，而实际出现的状态只聚集在状态空间的某些区域，因此，使用k-均值聚类对状态聚集的区域进行划分，可以在使用相同数目离散状态的条件下保证更高的离散化精度，从而提高优化的效果。

发明内容

[0004] 本发明的目的是使用k-均值聚类对交通状态进行离散化，来提高策略迭代的优化效果，更好地优化路口交通信号灯的控制策略。最终目的是为了增加单位时间内通过路口的车辆数，并且降低因等待红灯引起的停车次数和平均延误。

[0005] 本发明先使用感应控制方法对路口交通信号进行控制，每隔一段较短的单位时间间隔，路口机记录当前相位和下一相位的车辆排队长度、离开停车线的车辆数和交通信号控制器的控制动作。路口机采集到足够的样本后，对样本中的车辆排队长度进行k-均值聚类，得到离散交通状态。然后使用策略迭代对策略进行优化，并将优化好的策略保存在交通信号控制器中。每隔一段较短的单位时间间隔，路口机把检测到的当前相位和下一相位的车辆排队长度发送给交通信号控制器，交通信号控制器根据车辆排队长度和事先保存好的优化策略选择合适的相位动作，供路口机执行。

[0006] 本发明提出一种基于策略迭代和聚类的交通信号优化控制方法，包括以下步骤：

[0007] 步骤1，选择待优化的信号控制方案为固定相序控制，定义交通状态为当前相位和下一相位的车辆排队长度，定义控制动作为保持当前相位或切换到下一相位，定义直接回报为一个与单个采样周期内离开停车线的车辆数有关的变量，定义状态-动作对为离散交通状态和控制动作组成的数据向量，定义每个状态-动作对的Q值表示处于相应离散交通状态下采取控制动作后获得的期望累积回报，定义控制策略为每个离散交通状态应该执行的控制动作；

[0008] 步骤2，路口机把交通信号控制器的控制策略设置为感应控制，最小绿灯时间、最大绿灯时间设置为采样周期的正整数倍，单位绿灯延长时间与采样周期相同，路口机对交通状态、执行的相位动作和离开停车线的车辆数进行采样并记录样本，采样方法为：在每个

采样时刻记录交通状态、控制动作和每个采样周期离开停车线的车辆数；

[0009] 步骤3，路口机采集到指定数目的样本后，对样本中的交通状态进行离散化，离散化方法为：先对采样得到的交通状态进行归一化，并且去掉间距超过预设阈值的交通状态，再进行k-均值聚类，将得到的质心进行编号，每个质心对应一个离散交通状态，并且把归一化样本中的交通状态用最近的质心的编号表示，得到对应的离散交通状态；

[0010] 步骤4，路口机使用策略迭代优化策略，把优化得到的策略和步骤3中得到的质心保存在交通信号控制器中；

[0011] 步骤5，路口机设置交通信号控制器的控制策略为步骤4获得的控制策略，并把决策周期设置为采样周期，在每个决策时刻，交通信号控制器接收路口机检测到的交通状态，进行归一化，计算归一化后的交通状态到每个质心的距离，求出距离最近的质心，根据质心对应的离散交通状态查询控制策略，得到控制动作并发送至路口机执行。

[0012] 本发明较现有技术所具有的优点：

[0013] 在使用策略迭代优化交通信号控制策略之前，需要先对交通状态进行离散化——把两个相位的车辆排队长度构成的连续状态空间转化为离散状态空间，离散化的精度会影响策略迭代的优化效果。在不同的典型时段，实际的交通状态并非散布在整个状态空间，而是集中在某些区域。使用k-均值聚类算法得到的离散交通状态只考虑实际出现的交通状态集中的区域，而不像传统离散化方法那样把不存在实际交通状态的区域也考虑进去。因而，与传统方法相比，使用k-均值聚类算法后，使用相同数目的离散交通状态能够得到更高的离散化精度，从而提高策略迭代的优化效果。

附图说明

[0014] 图1为城市道路交叉口交通信号控制原理图。

[0015] 图2为一种基于策略迭代和聚类的交通信号优化控制方法流程图。

[0016] 1、第一地磁车辆检测器；2、第二地磁车辆检测器；3、第三地磁车辆检测器；4、第四地磁车辆检测器；5、第五地磁车辆检测器；6、第六地磁车辆检测器；7、第七地磁车辆检测器；8、第八地磁车辆检测器；9、第九地磁车辆检测器；10、第十地磁车辆检测器；11、第十一地磁车辆检测器；12、第十二地磁车辆检测器；13、第十三地磁车辆检测器；14、第十四地磁车辆检测器；15、第十五地磁车辆检测器；16、第十六地磁车辆检测器；17、第十七地磁车辆检测器；18、第十八地磁车辆检测器；19、第十九地磁车辆检测器；20、第二十地磁车辆检测器；21、第二十一地磁车辆检测器；22、第二十二地磁车辆检测器；23、第二十三地磁车辆检测器；24、第二十四地磁车辆检测器；25、车道一；26、车道二；27、车道三；28、车道四；29、车道五；30、车道六；31、车道七；32、车道八；33、车道九；34、车道十；35、车道十一；36、车道十二。

具体实施方式

[0017] 为使本发明的目的、技术方案和优点更加清楚，下面参照附图，对本发明作进一步详细说明。

[0018] 每个车道都需要安置两个地磁车辆检测器，一个地磁车辆检测器安置在停车线上游紧靠停车线处，检测通过停车线的车辆数，另一个地磁车辆检测器安置在停车线上游120

米处，检测通过停车线上游120米处断面的车辆数。通过这两个地磁车辆检测器可以计算其所在车道的任意时刻的位于停车线和停车线上游120米断面之间的车辆数，并换算成车辆排队长度。如图1所示，第一地磁车辆检测器1和第二地磁车辆检测器2用于检测车道一25的车辆排队长度，第三地磁车辆检测器3和第四地磁车辆检测器4用于检测车道二26的车辆排队长度，第五地磁车辆检测器5和第六地磁车辆检测器6用于检测车道三27的车辆排队长度，第七地磁车辆检测器7和第八地磁车辆检测器8用于检测车道四28的车辆排队长度，第九地磁车辆检测器9和第十地磁车辆检测器10用于检测车道五29的车辆排队长度，第十一地磁车辆检测器11和第十二地磁车辆检测器12用于检测车道六30的车辆排队长度，第十三地磁车辆检测器13和第十四地磁车辆检测器14用于检测车道七31的车辆排队长度，第十五地磁车辆检测器15和第十六地磁车辆检测器16用于检测车道八32的车辆排队长度，第十七地磁车辆检测器17和第十八地磁车辆检测器18用于检测车道九33的车辆排队长度，第十九地磁车辆检测器19和第二十地磁车辆检测器20用于检测车道十34的车辆排队长度，第二十一地磁车辆检测器21和第二十二地磁车辆检测器22用于检测车道十一35的车辆排队长度，第二十三地磁车辆检测器23和第二十四地磁车辆检测器24用于检测车道十二36的车辆排队长度。

[0019] 路口机接收第一地磁车辆检测器1至第二十四地磁车辆检测器24共计二十四个地磁车辆检测器发送的信息，然后转发至交通信号控制器。每隔10秒，交通信号控制器根据接收到的交通状态和路口机设置的控制策略决定控制动作。

[0020] 图2所示的一种基于策略迭代和聚类的交通信号优化控制方法流程图包含如下步骤：

[0021] 步骤1，选择信号控制方案，定义交通状态、控制动作、直接回报和Q值：

[0022] 待优化的信号控制方案采用固定相序控制方案，下面以四对称相位的情况为例介绍控制方案，但本发明不限于使用四相位、也不限于使用对称相位。相位1：允许车道一25和车道四28上的车辆直行和右转，允许车道二26和车道五29上的车辆直行；相位2：允许车道三27和车道六上30的车辆左转；相位3：允许车道七31和车道十34上的车辆直行和右转，允许车道八32和车道十一35上的车辆直行；相位4：允许车道九33和车道十二36上的车辆左转。交通信号在每个时刻只能处于四个相位中的一个，并且按照顺序依次执行。尽管相位顺序是固定的，每个相位的绿灯时长却不必固定。定义控制动作为保持当前相位或切换到下一相位，如果当前相位为相位1，则经过10秒后，交通信号控制器需要决策控制动作：保持相位1，或者切换到相位2，如果选择相位2，经过10秒又需要做出一次控制动作：保持相位2，或者切换到相位3，如果选择相位3，经过10秒又需要做出一次控制动作：保持相位3，或者切换到相位4，如果选择相位4，经过10秒又需要做出一次控制动作：保持相位4，或者切换到相位1……如此循环往复。定义所有相位的最小绿灯时间为10秒，最大绿灯时间为60秒。

[0023] 定义每个相位的车辆排队长度为该相位所有车道的车辆排队长度的最大值，相位1的车辆排队长度等于车道一25、车道二26、车道四28和车道五29的车辆排队长度中的最大值；相位2的车辆排队长度等于车道三27和车道六30的车辆排队长度中的最大值；相位3的车辆排队长度等于车道七31、车道八32、车道十34和车道十一35的车辆排队长度中的最大值；相位4的车辆排队长度等于车道九33和车道十二36的车辆排队长度中的最大值。

[0024] 定义交通状态为当前相位和下一相位的车辆排队长度，例如，如果当前相位为相

位1，则当前交通状态由相位1和相位2的车辆排队长度这两个变量组成的向量数据表示。

[0025] 定义一个采样周期的起始时刻为交通信号控制器决策控制动作的时刻，采样周期的时长与决策周期的时长相等，为10秒；定义直接回报为与单个采样周期离开停车线的车辆数有关的无量纲量，表示处于一个交通状态下采取控制动作后获得的直接好处；定义状态-动作对为离散交通状态和控制动作组成的数据向量；定义每个状态-动作对的Q值是处于相应离散交通状态下采取控制动作后获得的期望累积回报，即采取控制动作后数个采样周期内获得的直接回报之和的期望，Q值代表的是处于离散交通状态下采取控制动作后获得的长远利益；定义控制策略为给定离散交通状态时应该采取的控制动作；

[0026] 直接回报r的计算公式如下：

[0027] $$r=\begin{cases}\dfrac{n_p}{6.5}-1.0 & \text{当前相位为相位1或相位3}\\ \dfrac{n_p}{4.5}-1.0 & \text{当前相位为相位2或相位4}\end{cases}$$

[0028] 上式中，n_p表示一个采样周期内通过停车线的车辆数，公式中的常数6.5、4.5和-1.0的作用是使直接回报r维持在[-1,1]之间。交通信号控制器根据相邻两次路口机发送来的交通状态计算出n_p，然后按照上面的公式计算得到直接回报r。

[0029] 状态-动作对的Q值定义如下：

[0030] $$Q(s,a)=E(\sum_{k=1}^{k=T}\gamma^{k-1}r(s,a))$$

[0031] s表示离散交通状态，a表示在交通状态s下执行的控制动作，Q(s,a)表示状态-动作对s-a的Q值，E表示期望，r(s,a)表示在状态s下执行控制动作a获得的直接回报，γ是折扣因子，是一个介于0和1之间的实数，k表示遇到交通状态s后经历了第k个采样周期，经历交通状态s并执行控制动作a，经过一个采样周期后对应k＝1，T表示遇到交通状态s后采样终止于第T个采样周期，即累积回报的计算只使用T个采样周期的直接回报。

[0032] 步骤2，对交通状态、执行的控制动作和离开停车线的车辆数进行采样。

[0033] 在指定的典型时段，如早高峰或晚高峰时段进行一段时间的采样，在采样阶段，路口机把交通信号控制器的控制策略设置为感应控制，最小绿灯时间、最大绿灯时间和设置为采样周期的正整数倍，单位绿灯延长时间与采样周期相同，设置每个相位的最小绿灯时间为10秒，最大绿灯时间为60秒，单位绿灯延长时间为10秒。每一秒钟均按照下面的方法决策相位：当前相位绿灯时间小于10秒时，保持当前相位；当前相位绿灯时间超过或等于60秒时，切换到下一个相位；当前相位绿灯时间介于10秒和60秒时，当前相位有来车就延长绿灯时间10秒，没有来车就直接切换到下一相位。每隔10秒，路口机检测并存储下列信息作为样本：当前相位和下一相位的车辆排队长度、执行的控制动作和每个采样周期离开停车线的车辆数。设定要采集的样本数为9000。

[0034] 步骤3，路口机采集到9000个样本后，对样本中的交通状态进行离散化。把每个样本整理为数据向量(1,a,1',r)的形式，1表示某个采样时刻的交通状态，a表示交通状态为1时执行的控制动作，1'表示1之后下一个采样时刻的交通状态，r表示交通状态从1转移到1'的这个采样周期内获得的直接回报，可以使用原始样本中每个采样周期内离开停车线的车辆数，按照步骤1中直接回报r的计算公式计算得到。

[0035] 对样本中的交通状态进行预处理，先进行归一化，然后去掉间距超过预设阈值的交通状态。选择欧氏距离作为距离，设置阈值为0.1，先从样本中随机选择一个归一化的交通状态加入一个空的数据集，称为交通状态数据集，然后把样本中剩余的交通状态按照下列原则加入到数据集中：如果样本中的交通状态到交通状态数据集中所有交通状态的距离都大于0.1，则把该交通状态加入交通状态数据集，否则不加入。

[0036] 对交通状态数据集中的交通状态进行k-均值聚类，定义簇为相近交通状态的集合，每个簇对应一个离散交通状态，定义质心为簇包含的所有交通状态的质心，设置质心数为30，后开始聚类，步骤如下：

[0037] 步骤a，从交通状态数据集中随机选择30个不同的交通状态作为初始质心；

[0038] 步骤b，计算每个交通状态到每个质心的距离，将每个交通状态指派到最近的质心，形成30个簇；

[0039] 步骤c，重新计算每个簇的质心；

[0040] 步骤d，计算质心的变化量，即原先的质心和新的质心之间的距离，若所有簇的质心不再发生变化，k-均值聚类结束，否则执行步骤b。

[0041] k-均值聚类结束后，把每个样本(1,a,1',r)中的1和1'分别指派到最近的质心，即分别转化为离散交通状态s和s'，把样本整理为数据向量(s,a,s',r)。

[0042] 步骤4，在路口机中任意初始化一个交通信号控制策略，然后使用策略迭代方法优化策略，把优化得到的策略和步骤3中得到的质心保存在交通信号控制器中；

[0043] 在单路口交通信号控制优化问题中，共有30个离散交通状态，每个离散交通状态下都有两个控制动作——a_1表示保持当前相位，a_2表示切换到下一相位，策略的优化在路口机中进行，使用策略迭代方法进行优化，步骤如下：

[0044] 步骤a，设置迭代次数为1，初始化Q值和控制策略，计算状态转移矩阵和直接回报矩阵。把每个状态-动作对的Q值初始化为零，保存在矩阵Q中，根据样本(s,a,s',r)估算直接回报矩阵R_1和R_2，R_1,R_2分别保存执行控制动作a_1、a_2后获得的直接回报的期望，设i＝1,2,...,30,j＝1,2,...,30,k＝1,2,Q,R_1和R_2的定义分别如下：

[0045] $$Q=\begin{bmatrix} Q(s_1,a_1) \\ Q(s_1,a_2) \\ Q(s_2,a_1) \\ Q(s_2,a_2) \\ \vdots \\ Q(s_{30},a_1) \\ Q(s_{30},a_2) \end{bmatrix},\quad R_1=\begin{bmatrix} r(s_1,a_1,s_1) & r(s_1,a_1,s_2) & \cdots & r(s_1,a_1,s_{30}) \\ r(s_2,a_1,s_1) & r(s_2,a_1,s_2) & \cdots & r(s_2,a_1,s_{30}) \\ \vdots & \vdots & & \vdots \\ r(s_{30},a_1,s_1) & r(s_{30},a_1,s_2) & \cdots & r(s_{30},a_1,s_{30}) \end{bmatrix},$$

[0046] $$R_2=\begin{bmatrix} r(s_1,a_2,s_1) & r(s_1,a_2,s_2) & \cdots & r(s_1,a_2,s_{30}) \\ r(s_2,a_2,s_1) & r(s_2,a_2,s_2) & \cdots & r(s_2,a_2,s_{30}) \\ \vdots & \vdots & & \vdots \\ r(s_{30},a_2,s_1) & r(s_{30},a_2,s_2) & \cdots & r(s_{30},a_2,s_{30}) \end{bmatrix}。$$

[0047] 其中，Q(s_i,a_k)表示动作-状态对s_i-a_k的Q值，r(s_i,a_k,s_j)表示处于离散交通状态s_i，执行控制动作a_k之后，转移到离散交通状态s_j时获得的直接回报。初始化一个控制策略

为任意策略，保存在矩阵Π中，Π的定义如下：

[0048]

$$\Pi=\begin{bmatrix}\pi(s_1,a_1) & \pi(s_1,a_2) & 0 & 0 & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & \pi(s_2,a_1) & \pi(s_2,a_2) & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & 0 & 0 & \pi(s_3,a_1) & \pi(s_3,a_2) & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \pi(s_{30},a_1) & \pi(s_{30},a_2)\end{bmatrix}。$$

[0049] 其中，$\pi(s_i,a_k)$表示在离散状态s_i下执行动作a_k的概率，Π的每行元素之和为1。根据样本(s,a,s',r)估算状态转移矩阵P，定义如下：

[0050]

$$P=\begin{bmatrix}p(s_1|s_1,a_1) & p(s_2|s_1,a_1) & \cdots & p(s_{30}|s_1,a_1)\\ p(s_1|s_1,a_2) & p(s_2|s_1,a_2) & \cdots & p(s_{30}|s_1,a_2)\\ p(s_1|s_2,a_1) & p(s_2|s_2,a_1) & \cdots & p(s_{30}|s_2,a_1)\\ p(s_1|s_2,a_2) & p(s_2|s_2,a_2) & \cdots & p(s_{30}|s_2,a_2)\\ \vdots & \vdots & & \vdots\\ p(s_1|s_{30},a_1) & p(s_2|s_{30},a_1) & \cdots & p(s_{30}|s_{30},a_1)\\ p(s_1|s_{30},a_2) & p(s_2|s_{30},a_2) & \cdots & p(s_{30}|s_{30},a_2)\end{bmatrix}$$

[0051] 其中，矩阵元素$p(s_j|s_i,a_k)$是条件概率，表示处于离散交通状态s_i，执行控制动作a_k之后，下一个采样时刻转移到离散交通状态s_j的概率。利用R_1，R_2和P中的元素，可以求出直接回报矩阵R，R的定义如下：

[0052]

$$R=\begin{bmatrix}r(s_1,a_1)\\ r(s_1,a_2)\\ r(s_2,a_1)\\ r(s_2,a_2)\\ \vdots\\ r(s_{30},a_1)\\ r(s_{30},a_2)\end{bmatrix}。$$

[0053] 其中，$r(s_i,a_k)$表示处于离散交通状态s_i，执行控制动作a_k之后获得的直接回报的期望，计算公式如下：

[0054]

$$r(s_i,a_k)=\sum_{j=1}^{30}r(s_i,a_k,s_j)p(s_j|s_i,a_k)。$$

[0055] 步骤b，更新Q值，按照下式更新矩阵Q：

[0056] $Q=(I-\gamma P\Pi)^{-1}R$

[0057] 其中，I表示单位矩阵，γ是折扣因子，设置为0.95，$()^{-1}$表示对矩阵求逆；

[0058] 步骤c，根据Q值更新控制策略，按照下式更新矩阵Π中的元素：

[0059]

$$\pi(s_i,a_k)=\begin{cases}1 & a_k=\arg\max\limits_{a\in\{a_1,a_2\}}Q(s_i,a)\\ 0 & a_k\neq\arg\max\limits_{a\in\{a_1,a_2\}}Q(s_i,a)\end{cases}$$

[0060] 步骤d，如果迭代次数为1，保存矩阵Π到一个同维矩阵Π'，迭代次数加1，返回步

骤b，否则，求解矩阵Π与矩阵Π'的差的二范数：

[0061] $D=||\Pi-\Pi'||$

[0062] 如果D等于0，则策略迭代结束，如果D不等于0，保存矩阵Π到矩阵Π'，迭代次数加1，返回步骤b。

[0063] 策略迭代结束后，得到的控制策略保存在矩阵Q中，把矩阵Q和步骤3中得到的质心保存在交通信号控制器中；

[0064] 步骤5，路口机设置交通信号控制器的控制策略为步骤4获得的控制策略，每隔10秒钟，交通信号控制器接收路口机检测到的交通状态，对其进行归一化，计算归一化后的交通状态到每个质心的距离，求出距离最近的质心的编号，即离散交通状态s_i状态的编号i，然后根据下式选择控制动作a^*：

[0065] $$a^* = \arg\max_{a\in\{a_1,a_2\}} Q(s_i,a))$$

[0066] 交通信号控制器把控制动作a^*发送至路口机执行，如果a^*的值为a_1则保持当前相位，如果a^*的值为a_2则切换到下一相位。

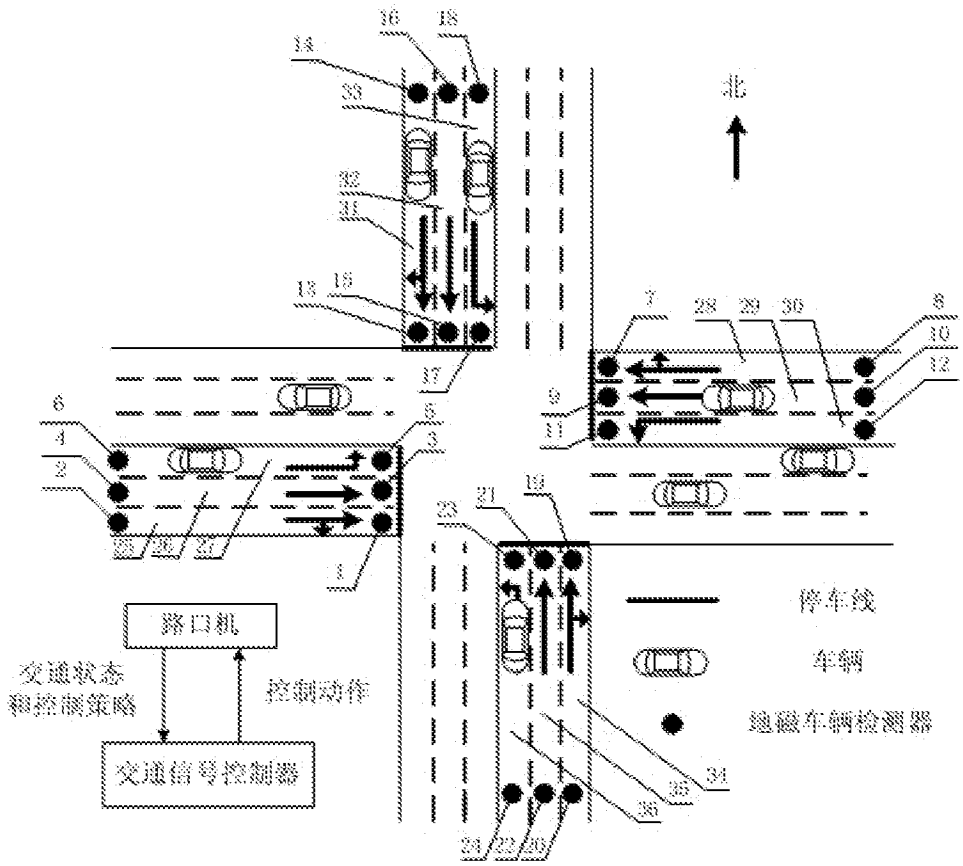


图1

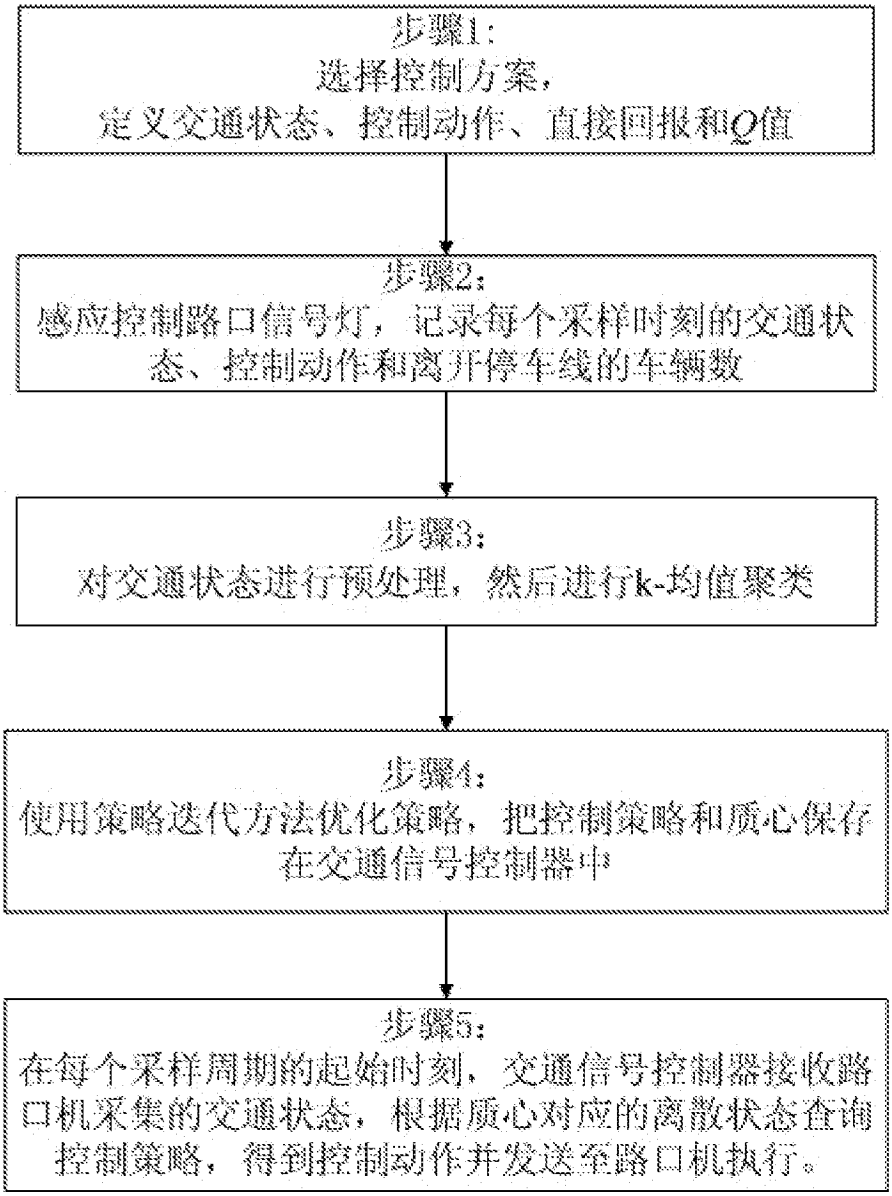

步骤1：
选择控制方案，
定义交通状态、控制动作、直接回报和Q值
步骤2：
感应控制路口信号灯，记录每个采样时刻的交通状态、控制动作和离开停车线的车辆数
步骤3：
对交通状态进行预处理，然后进行k-均值聚类
步骤4：
使用策略迭代方法优化策略，把控制策略和质心保存在交通信号控制器中
步骤5：
在每个采样周期的起始时刻，交通信号控制器接收路口机采集的交通状态，根据质心对应的离散状态查询控制策略，得到控制动作并发送至路口机执行。

图2