

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2018 年 11 月 15 日 (15.11.2018)



(10) 国际公布号
WO 2018/205245 A1

(51) 国际专利分类号:
G06N 5/02 (2006.01) **G06N 99/00** (2010.01)

中国广东省深圳市南山区西丽深圳大学城学苑大道1068号, Guangdong 518060 (CN)。

(21) 国际申请号: PCT/CN2017/084081

(22) 国际申请日: 2017 年 5 月 12 日 (12.05.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 中国科学院深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES) [CN/CN]; 中国广东省深圳市南山区西丽深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(74) 代理人: 深圳青年人专利商标代理有限公司 (SHENZHEN YOUTH PATENT AND TRADEMARK AGENCY LTD.); 中国广东省深圳市罗湖区深南东路 5045 号深业中心大厦 2502-2503, Guangdong 518010 (CN)。

(72) 发明人: 李慧云(LI, Huiyun); 中国广东省深圳市南山区西丽深圳大学城学苑大道1068号, Guangdong 518060 (CN)。 王峥(WANG, Zheng); 中国广东省深圳市南山区西丽深圳大学城学苑大道1068号, Guangdong 518060 (CN)。 刘玢玢(LIU, Fenfen);

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,

(54) Title: STRATEGY NETWORK MODEL GENERATION METHOD AND APPARATUS FOR AUTOMATIC VEHICLE DRIVING

(54) 发明名称: 一种用于车辆自动驾驶的策略网络模型的生成方法及装置

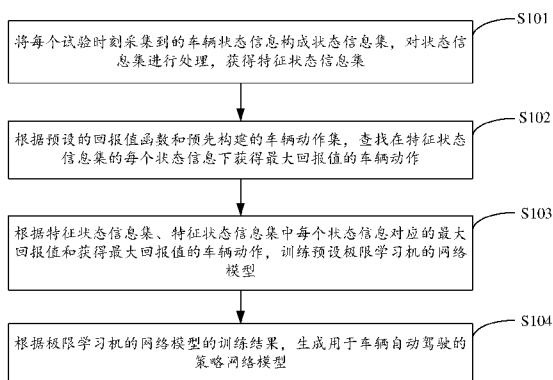


图 1

- S101 Form vehicle state information collected at each test moment into a state information set, and process the state information set, so as to obtain a feature state information set
- S102 Search, according to a pre-set reward value function and a pre-constructed vehicle action set, for a vehicle action obtaining the maximum reward value from each piece of state information in the feature state information set
- S103 Train a network model of an extreme learning machine according to the feature state information set, the maximum reward value corresponding to each piece of state information in the feature state information set, and the vehicle action obtaining the maximum reward value
- S104 Generate, according to a training result of the network model of the extreme learning machine, a strategy network model for automatic vehicle driving

(57) Abstract: A strategy network model generation method and apparatus for automatic vehicle driving. The method comprises: forming vehicle state information collected at each test moment into a state information set, and processing the state information set, so as to obtain a feature state information set (S101); searching, according to a pre-set reward value function and a pre-constructed vehicle action set, for a vehicle action obtaining the maximum reward value from each piece of state information in the feature state information set (S102); training a network model of an extreme learning machine according to the feature state information set, the maximum reward value corresponding to each piece of state information in the feature state information set, and the vehicle action obtaining the maximum reward value (S103); and generating, according to a training result of the network model of the extreme learning machine, a strategy network model for automatic vehicle driving (S104), thereby effectively reducing the consumption of computation resources, and effectively improving the generation efficiency of the strategy network model for automatic vehicle driving.

NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：提供了一种用于车辆自动驾驶的策略网络模型的生成方法及装置，所述方法包括：将每个试验时刻采集到的车辆状态信息构成状态信息集，对状态信息集进行处理，获得特征状态信息集(S101)；根据预设的回报值函数和预先构建的车辆动作集，查找在特征状态信息集的每个状态信息下获得最大回报值的车辆动作(S102)；根据特征状态信息集、特征状态信息集中每个状态信息对应的最大回报值和获得最大回报值的车辆动作，训练极限学习机的网络模型(S103)；根据极限学习机的网络模型的训练结果，生成用于车辆自动驾驶的策略网络模型(S104)，从而有效地降低了计算资源的消耗，有效地提高了车辆自动驾驶策略网络模型的生成效率。

一种用于车辆自动驾驶的策略网络模型的生成方法及装置

技术领域

5 本发明属于计算机技术领域，尤其涉及一种用于车辆自动驾驶的策略网络模型的生成方法及装置。

背景技术

随着经济的发展和城镇化的推进，全球汽车保有量和道路里程逐步增加，
10 导致诸如交通拥堵、事故、污染、土地资源紧缺等一系列传统汽车无法妥善解决的问题日益凸显。无人驾驶汽车技术被视为这些问题的有效解决方案，其发展被受瞩目，美国电子和电子工程师协会（IEEE）预测，至 2040 年自动驾驶车辆数目所占比例将达到 75%。

无人驾驶汽车，即在没有司机的情况下通过自身的辅助驾驶系统在道路上
15 行驶，具备环境感知能力。辅助驾驶系统通过传感器获取的环境信息做出控制决策，已出现了多种用于辅助驾驶系统的传感器和产品，例如：激光雷达、机器视觉、自适应巡航、车辆接近通报装置、夜视辅助、自适应前照明系统等。

目前，辅助驾驶系统的控制方法主要为基于规则的控制决策，即根据已知的驾驶经验构建对车辆环境信息输出控制决策的专家规则系统，专家规则系统
20 这类浅层学习算法可看作从被标记的数据之间寻找规律的过程，当规则很难被抽象成公式或简单逻辑时，浅层学习就无法奏效，然而，自动驾驶场景类别多样、路况复杂，很难用有限的规则定义清楚。

随着深度强化学习技术的快速发展，一些研究机构提出“端对端”式的自动驾驶算法，通过深度网络构建辅助驾驶系统中的控制决策模型。这类方法不需
25 要对车辆的状态进行基于规则的识别。然而，对车辆自动驾驶进行深度学习要求强大的计算资源，例如需要数百瓦的图形处理器（GPU）等长时间（数天到

数周)的训练,才能得到一个决策网络,如果外界条件改变,则需要重新训练。

发明内容

5 本发明的目的在于提供一种用于车辆自动驾驶的策略网络模型的生成方法及装置,旨在解决现有技术中策略网络模型的生成效率较低的问题。

一方面,本发明提供了一种用于车辆自动驾驶的策略网络模型的生成方法,所述方法包括下述步骤:

将每个试验时刻采集到的车辆状态信息构成状态信息集,对所述状态信息集进行处理,获得特征状态信息集;

10 根据预设的回报值函数和预先构建的车辆动作集,查找在所述特征状态信息集的每个状态信息下获得最大回报值的车辆动作;

根据所述特征状态信息集、所述特征状态信息集中每个状态信息对应的最大回报值和所述获得最大回报值的车辆动作,训练极限学习机的网络模型;

15 根据所述极限学习机的网络模型的训练结果,生成用于所述车辆自动驾驶的策略网络模型。

另一方面,本发明提供了一种用于车辆自动驾驶的策略网络模型的生成装置,所述装置包括:

状态处理模块,用于将每个试验时刻采集到的车辆状态信息构成状态信息集,对所述状态信息集进行处理,获得特征状态信息集;

20 强化学习模块,用于根据预设的回报值函数和预先构建的车辆动作集,查找在所述特征状态信息集的每个状态信息下获得最大回报值的车辆动作;

极限学习机训练模块,用于根据所述特征状态信息集、所述特征状态信息集中每个状态信息对应的最大回报值和所述获得最大回报值的车辆动作,训练极限学习机的网络模型;以及

25 策略网络生成模块,用于根据所述极限学习机的网络模型的训练结果,生成用于所述车辆自动驾驶的策略网络模型。

本发明将每个试验时刻采集到的车辆状态信息构成状态信息集，对状态信息集进行处理，以在降低状态信息集中数据量的同时保留状态信息集中的特征，获得特征状态信息集，根据预设的回报值函数和预先构建的车辆动作集，查找在特征状态信息集的每个状态信息下获得最大回报值的车辆动作，根据特征状态信息集中的状态信息、每个状态信息对应的最大回报值以及每个最大回报值对应的车辆动作，训练预设的极限学习机的网络模型，并根据训练结果，生成车辆自动驾驶的策略模型，从而通过对状态信息的处理、通过采用强化学习得到每个状态信息的最大回报值和最大回报值对应的车辆动作、通过训练极限学习及的网络模型，有效地降低了计算资源的消耗，有效地提高了车辆自动驾驶策略网络模型的泛化能力和生成效率。

附图说明

图 1 是本发明实施例一提供的用于车辆自动驾驶的策略网络模型的生成方法的实现流程图；

图 2 是本发明实施例一提供的用于车辆自动驾驶的策略网络模型的结构示例图；

图 3 是本发明实施例二提供的用于车辆自动驾驶的策略网络模型的生成装置的结构示意图；以及

图 4 是本发明实施例二提供的用于车辆自动驾驶的策略网络模型的生成装置的优选结构示意图。

具体实施方式

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。

以下结合具体实施例对本发明的具体实现进行详细描述：

实施例一：

图 1 示出了本发明实施例一提供的用于车辆自动驾驶的策略网络模型的生成方法的实现流程，为了便于说明，仅示出了与本发明实施例相关的部分，详述如下：

- 5 在步骤 S101 中，将每个试验时刻采集到的车辆状态信息构成状态信息集，对状态信息集进行处理，获得特征状态信息集。

本发明适用于基于赛车仿真模拟平台或赛车模拟器（例如开放式赛车模拟器 TORCS, The open racing car simulation）建立的交互平台，在该交互平台上进行无人驾驶汽车的行驶交互试验。在当前的交互试验过程中，通过车辆上预
10 设的多个传感器采集车辆状态信息，车辆状态信息可包括车辆离道路中线的距离、车辆前进方向与道路切向的夹角、车辆前方激光测距仪的距离值和车辆在道路切向上的速度分量。

在本发明实施例中，采集每个实验时刻的车辆状态信息，并将所有的车辆状态信息一起构成状态信息集，为了减少状态信息集中的数据量并保留状态信息
15 集的特征，对状态信息集进行处理，生成特征状态信息集，从而通过对状态信息集的处理有效地降低后续过程的计算量。其中，特征状态信息集可包含具有代表性的状态信息或状态信息的特征。

具体地，可通过预设的聚类算法（例如 K-means 聚类）将状态信息集划分为若干个簇，并聚类出能够代表每个簇的簇中心点，这些簇中心点的状态信息
20 构成特征状态信息集，此外，还可通过预设的特征提取算法（例如自编码器特征提取算法）提取状态信息集中的特征，由这些特征构成特征状态信息集。

在步骤 S102 中，根据预设的回报值函数和预先构建的车辆动作集，查找在特征状态信息集的每个状态信息下获得最大回报值的车辆动作。

在本发明实施例中，回报值函数用来计算在某个状态下车辆做出某个动作
25 后获得的立即回报值，车辆动作集中可包括直行、刹车等车辆动作。先对特征状态信息集的每个状态信息对应的回报值序列进行初始化，可初始化为零，每

个状态信息的回报值序列包括在该状态信息下每个车辆动作对应的回报值，其中，状态信息 s 的回报值序列中车辆动作 s 的回报值可表示为 $\hat{Q}(s,a)$ 。

在本发明实施例中，根据回报值函数和车辆动作集，对特征状态信息集中状态信息的回报值序列进行多次更新，直至所有回报值序列中的值不再变化，最后从所有回报值序列中选取特征状态信息集中每个状态信息的最大回报值以及该最大回报值对应的车辆动作，从而通过强化学习确定了特征状态信息集中每个状态信息最佳的车辆动作。

具体地，在更新回报值序列时，在车辆动作集中遍历可以在特征状态信息集的当前状态信息下执行的车辆动作，执行遍历得到的动作，并通过回报值函数计算执行后的立即回报值，根据该立即回报值和预设的更新公式，即可对在该当前状态信息下遍历到的车辆动作所对应的回报值。其中，回报值函数可表示为：

$\hat{Q}(s,a) = r + \gamma \cdot \max(\hat{Q}(s',a'))$ ，其中， $\hat{Q}(s,a)$ 为在特征状态信息集的状态信息 s 下执行遍历到的动作 a 时的回报值， $\max(\hat{Q}(s',a'))$ 为状态信息 s 的下一状态信息 s' 的最大回报值， r 为在状态信息 s 下执行遍历到的动作 a 时的立即回报值， γ 为预设参数且 $0 \leq \gamma < 1$ 。

作为示例地，回报值函数的公式可为：

$r = \Delta dis * \cos(\alpha * angle) * \text{sgn}(trackPos - threshold)$ ，其中， r 为回报值函数计算得到的立即回报值， Δdis 为车辆在相邻试验时刻跑过的有效距离， α 为预设的权重缩放因子， $angle$ 为车辆当前行驶方向与道路切向的夹角， $trackPos$ 为车辆离道路中线的距离， $threshold$ 为预设阈值，当 $trackPos$ 大于 $threshold$ 时， r 为无穷小，可表示对车辆太靠近道路边界时的惩罚。另外，回报值函数还可考虑行驶速度、单位耗油量、平稳度等。

在步骤 S103 中，根据特征状态信息集、特征状态信息集中每个状态信息对应的最大回报值和获得最大回报值的车辆动作，训练预设极限学习机的网络模型。

在本发明实施例中，极限学习机是一种单隐层前馈神经网络学习算法。其中，极限学习机的网络模型可表示为：

$$\sum_{i=1}^M \beta_i g(w_i \cdot x_i + b_i) = Q_j, \text{ 其中, } g(\cdot) \text{ 为隐层节点的激活函数, } \beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{in}]^T$$

为连接第 i 个隐层节点与网络输出层节点的输出权值向量， $\omega_i = [\omega_{i1}, \omega_{i2}, \dots, \omega_{mi}]$ 为连接第 i 个隐层节点的随机输入权值向量，极限学习机网络模型的输出值为 $Q_i = [Q_{i1}, Q_{i2}, \dots, Q_{in}]^T$ ， Q_j 为输出值中的第 j 列向量， $j=1, 2, \dots, N$ ， N 为输出节点的个数， M 为隐层节点的个数。

在本发明实施例中，可将特征状态信息集中每个状态信息对应的最大回报值设置为极限学习机网络模型一次训练过程中的输出值，将特征状态信息集中的状态信息和每个状态信息下获得最大回报值的车辆动作最为极限学习机网络模型一次训练过程中的输入值，从而训练得到极限学习机网络模型中隐层节点与网络输出层节点的输出权值向量。可通过进行多次的车辆自动驾驶试验，多次地训练极限学习机网络模型。

在本发明实施例中，极限学习机网络模型的表示方式还可写为矩阵形式 $W\beta = Q$ ，当隐层节点的激活函数无限可微时，极限学习机网络模型的输入权值和隐层节点阈值可随机赋值，此时可将矩阵 W 视为常数，极限学习机的学习过程可等价于求取线性系统 $W\beta = Q$ 最小范数的最小二乘解。

在步骤 S104 中，根据极限学习机的网络模型的训练结果，生成用于车辆自动驾驶的策略网络模型。

在本发明实施例中，图 2 为策略网络模型的示例，由此可见，极限学习机网络模型中隐层节点与网络输出层节点的输出权值向量，即策略网络模型中隐藏层的权重值，最后得到了状态数据处理、强化学习和极限学习机网络模型构成的策略网络模型。

在本发明实施例中，对采集到的所有状态信息构成的状态信息集进行处理，以降低状态信息集中的数据量并保留状态信息集中的特征，生成特征状态信息

集，通过强化学习得到特征状态信息集中每个状态信息的最大回报值和该最大回报值对应的车辆动作，通过特征状态信息集中每个状态信息、每个状态信息的最大回报值、和最大回报值对应的车辆动作，训练极限学习机的网络模型，最后根据训练的到的参数向量，得到车辆自动驾驶的策略网络模型，从而通过对状态信息的处理、通过强化学习和极限学习机的结合，有效地降低了计算资源的消耗，有效地提高了车辆自动驾驶策略网络模型的泛化能力和生成效率。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，所述的程序可以存储于一计算机可读取存储介质中，所述的存储介质，如 ROM/RAM、磁盘、光盘等。

10 实施例二：

图 3 示出了本发明实施例二提供的用于车辆自动驾驶的策略网络模型的生成装置的结构，为了便于说明，仅示出了与本发明实施例相关的部分，其中包括：

状态处理模块 31，用于将每个试验时刻采集到的车辆状态信息构成状态信息集，对状态信息集进行处理，获得特征状态信息集。

在本发明实施例中，采集每个实验时刻的车辆状态信息，并将所有的车辆状态信息一起构成状态信息集，为了减少状态信息集中的数据量并保留状态信息集的特征，对状态信息集进行处理，生成特征状态信息集，从而通过对状态信息集的处理有效地降低后续过程的计算量。其中，特征状态信息集可包含具有代表性的状态信息或状态信息的特征。

强化学习模块 32，用于根据预设的回报值函数和预先构建的车辆动作集，查找在特征状态信息集的每个状态信息下获得最大回报值的车辆动作。

在本发明实施例中，每个状态信息的回报值序列包括在该状态信息下每个车辆动作对应的回报值，根据回报值函数和车辆动作集，对特征状态信息集中状态信息的回报值序列进行多次更新，直至所有回报值序列中的值不再变化，最后从所有回报值序列中选取特征状态信息集中每个状态信息的最大回报值以

及该最大回报值对应的车辆动作，从而通过强化学习确定了特征状态信息集中每个状态信息最佳的车辆动作。

具体地，在更新回报值序列时，在车辆动作集中遍历可以在特征状态信息集的当前状态信息下执行的车辆动作，执行遍历得到的动作，并通过回报值函数计算执行后的立即回报值，根据该立即回报值和预设的更新公式，即可对在该当前状态信息下遍历到的车辆动作所对应的回报值。其中，回报值函数可表示为：

$\hat{Q}(s,a)=r+\gamma\cdot\max(\hat{Q}(s',a'))$ ，其中， $\hat{Q}(s,a)$ 为在特征状态信息集的状态信息 s 下执行遍历到的动作 a 时的回报值， $\max(\hat{Q}(s',a'))$ 为状态信息 s 的下一状态信息 s' 的最大回报值， r 为在状态信息 s 下执行遍历到的动作 a 时的立即回报值， γ 为预设参数且 $0\leq\gamma<1$ 。

极限学习机训练模块 33，用于根据特征状态信息集、特征状态信息集中每个状态信息对应的最大回报值和获得最大回报值的车辆动作，训练极限学习机的网络模型。

在本发明实施例中，极限学习机的网络模型可表示为：

$$\sum_{i=1}^M \beta_i g(w_i \cdot x_i + b_i) = Q_j, \text{ 其中, } g(\cdot) \text{ 为隐层节点的激活函数, } \beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{in}]^T$$

为连接第 i 个隐层节点与网络输出层节点的输出权值向量， $\omega_i = [\omega_{i1}, \omega_{i2}, \dots, \omega_{mi}]$ 为连接第 i 个隐层节点的随机输入权值向量，极限学习机网络模型的输出值为 $Q_i = [Q_{i1}, Q_{i2}, \dots, Q_{in}]^T$ ， Q_j 为输出值中的第 j 列向量， $j=1,2,\dots,N$ ， N 为输出节点的个数， M 为隐层节点的个数。

在本发明实施例中，可将特征状态信息集中每个状态信息对应的最大回报值设置为极限学习机网络模型一次训练过程中的输出值，将特征状态信息集中的状态信息和每个状态信息下获得最大回报值的车辆动作最为极限学习机网络模型一次训练过程中的输入值，从而训练得到极限学习机网络模型中隐层节点与网络输出层节点的输出权值向量。可通过进行多次的车辆自动驾驶试验，多

次地训练极限学习机网络模型。

在本发明实施例中，极限学习机网络模型的表示方式还可写为矩阵形式 $W\beta=Q$ ，当隐层节点的激活函数无限可微时，极限学习机网络模型的输入权值和隐层节点阈值可随机赋值，此时可将矩阵 W 视为常数，极限学习机的学习过程可等价于求取线性系统 $W\beta=Q$ 最小范数的最小二乘解。

策略网络生成模块 34，用于根据极限学习机的网络模型的训练结果，生成用于车辆自动驾驶的策略网络模型。

在本发明实施例中，极限学习机网络模型中隐层节点与网络输出层节点的输出权值向量，即策略网络模型中隐藏层的权重值，最后得到了状态数据处理、强化学习和极限学习机网络模型构成的策略网络模型。

优选地，如图 4 所示，状态处理模块 31 包括：

状态处理子模块 411，用于通过预设的聚类算法对状态信息集进行聚类、或通过预设的特征提取算法对状态信息集进行特征提取，以生成特征状态信息集。

优选地，强化学习模块 32 包括回报值初始化模块 421、回报值更新模块 422 和最大回报值获取模块 423，其中：

回报值初始化模块 421，用于对特征状态信息集的每个状态信息对应的回报值序列进行初始化，回报值序列包括执行车辆动作集中所有车辆动作对应的回报值；

回报值更新模块 422，用于根据回报值函数和车辆动作集，对所有回报值序列进行多次更新，直至所有回报值序列不再变化；以及

最大回报值获取模块 423，用于从特征状态信息的每个状态信息对应的回报值序列中获取最大回报值和最大回报值对应的车辆动作。

回报值更新模块 422 包括动作遍历模块和回报值更新子模块，其中：

动作遍历模块，用于遍历车辆动作集，并根据回报值函数计算在特征状态信息集的状态信息下执行遍历到的车辆动作时的立即回报值；以及

回报值更新子模块，用于根据立即回报值和预设的更新公式，更新在特征状态信息集的状态信息下执行遍历到的车辆动作对应的回报值。

优选地，极限学习机训练模块 33 包括学习机设置模块 431 和网络训练模块 432，其中：

- 5 学习机设置模块 431，用于根据特征状态信息集、特征状态信息集中每个状态信息对应的最大回报值和获得最大回报值的车辆动作，设置极限学习机网络模型的输入和输出；以及

网络训练模块 432，用于对极限学习机网络模型进行训练，获得极限学习机网络模型的输出权重。

- 10 在本发明实施例中，对采集到的所有状态信息构成的状态信息集进行处理，以降低状态信息集中的数据量并保留状态信息集中的特征，生成特征状态信息集，通过强化学习得到特征状态信息集中每个状态信息的最大回报值和该最大回报值对应的车辆动作，通过特征状态信息集中每个状态信息、每个状态信息的最大回报值、和最大回报值对应的车辆动作，训练极限学习机的网络模型，
- 15 最后根据训练的到的参数向量，得到车辆自动驾驶的策略网络模型，从而通过对状态信息的处理、通过强化学习和极限学习机的结合有效地降低了计算资源的消耗，有效地提高了车辆自动驾驶策略网络模型的泛化能力和生成效率。

- 20 在本发明实施例中，用于车辆自动驾驶的策略网络模型的装置各模块可由相应的硬件或软件模块实现，各模块可以为独立的软、硬件模块，也可以集成为一个软、硬件模块，在此不用以限制本发明。

以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1、一种用于车辆自动驾驶的策略网络模型的生成方法，其特征在于，所述方法包括下述步骤：

将每个试验时刻采集到的车辆状态信息构成状态信息集，对所述状态信息
5 集进行处理，获得特征状态信息集；

根据预设的回报值函数和预先构建的车辆动作集，查找在所述特征状态信息集的每个状态信息下获得最大回报值的车辆动作；

根据所述特征状态信息集、所述特征状态信息集中每个状态信息对应的最大回报值和所述获得最大回报值的车辆动作，训练预设极限学习机的网络模型；

10 根据所述极限学习机的网络模型的训练结果，生成用于所述车辆自动驾驶的策略网络模型。

2、如权利要求 1 所述的方法，其特征在于，将每个试验时刻采集到的车辆状态信息构成状态信息集，对所述状态信息集进行处理，获得特征状态信息集的步骤，包括：

15 通过预设的聚类算法对所述状态信息集进行聚类、或通过预设的特征提取算法对所述状态信息集进行特征提取，以生成所述特征状态信息集。

3、如权利要求 1 所述的方法，其特征在于，根据预设的回报值函数和预先构建的车辆动作集，查找在所述特征状态信息集的每个状态信息下获得最大回报值的车辆动作的步骤，包括：

20 对所述特征状态信息集的每个状态信息对应的回报值序列进行初始化，所述回报值序列包括执行所述车辆动作集中所有车辆动作对应的回报值；

根据所述回报值函数和所述车辆动作集，对所述所有回报值序列进行多次更新，直至所述所有回报值序列不再变化；

25 从所述特征状态信息的每个状态信息对应的回报值序列中获取最大回报值和所述最大回报值对应的车辆动作。

4、如权利要求 3 所述的方法，其特征在于，根据所述回报值函数和所述车

辆动作集，对所述所有回报值序列进行多次更新，直至所述所有回报值序列不再变化的步骤，包括：

遍历所述车辆动作集，并根据所述回报值函数计算在所述特征状态信息集的状态信息下执行遍历到的车辆动作时的立即回报值；

- 5 根据所述立即回报值和预设的更新公式，更新在所述特征状态信息集的状态信息下执行所述遍历到的车辆动作对应的回报值，所述更新公式为：

$$\hat{Q}(s,a)=r+\gamma\cdot\max(\hat{Q}(s',a'))$$
，其中， $\hat{Q}(s,a)$ 为在所述特征状态信息集的状态信息 s 下执行所述遍历到的动作 a 时的回报值， $\max(\hat{Q}(s',a'))$ 为所述状态信息 s 的下一状态信息 s' 的最大回报值， r 为在所述状态信息 s 下执行所述遍历到的动作 a 时的立即回报值， γ 为预设参数且 $0\leq\gamma<1$ 。

10

5、如权利要求 1 所述的方法，其特征在于，根据所述特征状态信息集、所述特征状态信息集中每个状态信息对应的最大回报值和所述获得最大回报值的车辆动作，训练预设极限学习机的网络模型的步骤，包括：

根据所述特征状态信息集、所述特征状态信息集中每个状态信息对应的最大回报值和所述获得最大回报值的车辆动作，设置所述极限学习机网络模型的输入和输出；

15

对所述极限学习机网络模型进行训练，获得所述极限学习机网络模型的输出权重。

- 6、一种用于车辆自动驾驶的策略网络模型的生成装置，其特征在于，所述装置包括：
- 20

状态处理模块，用于将每个试验时刻采集到的车辆状态信息构成状态信息集，对所述状态信息集进行处理，获得特征状态信息集；

强化学习模块，用于根据预设的回报值函数和预先构建的车辆动作集，查找在所述特征状态信息集的每个状态信息下获得最大回报值的车辆动作；

- 25 极限学习机训练模块，用于根据所述特征状态信息集、所述特征状态信息集中每个状态信息对应的最大回报值和所述获得最大回报值的车辆动作，训练

极限学习机的网络模型；以及

策略网络生成模块，用于根据所述极限学习机的网络模型的训练结果，生成用于所述车辆自动驾驶的策略网络模型。

7、如权利要求6所述的装置，其特征在于，所述状态处理模块包括：

5 状态处理子模块，用于通过预设的聚类算法对所述状态信息集进行聚类、或通过预设的特征提取算法对所述状态信息集进行特征提取，以生成所述特征状态信息集。

8、如权利要求6所述的装置，其特征在于，所述强化学习模块包括：

10 回报值初始化模块，用于对所述特征状态信息集的每个状态信息对应的回报值序列进行初始化，所述回报值序列包括执行所述车辆动作集中所有车辆动作对应的回报值；

回报值更新模块，用于根据所述回报值函数和所述车辆动作集，对所述所有回报值序列进行多次更新，直至所述所有回报值序列不再变化；以及

15 最大回报值获取模块，用于从所述特征状态信息的每个状态信息对应的回报值序列中获取最大回报值和所述最大回报值对应的车辆动作。

9、如权利要求8所述的装置，其特征在于，所述回报值更新模块包括：

动作遍历模块，用于遍历所述车辆动作集，并根据所述回报值函数计算在所述特征状态信息集的状态信息下执行遍历到的车辆动作时的立即回报值；以及

20 回报值更新子模块，用于根据所述立即回报值和预设的更新公式，更新在所述特征状态信息集的状态信息下执行所述遍历到的车辆动作对应的回报值，所述更新公式为：

$$\hat{Q}(s,a)=r+\gamma\cdot\max(\hat{Q}(s',a'))$$
，其中， $\hat{Q}(s,a)$ 为在所述特征状态信息集的状态信息 s 下执行所述遍历到的动作 a 时的回报值， $\max(\hat{Q}(s',a'))$ 为所述状态信息 s 的下一状态信息 s' 的最大回报值， r 为在所述状态信息 s 下执行所述遍历到的动作 a 时的立即回报值， γ 为预设参数且 $0\leq\gamma<1$ 。

25

10、如权利要求 6 所述的装置，其特征在于，所述极限学习机训练模块包括：

学习机设置模块，用于根据所述特征状态信息集、所述特征状态信息集中每个状态信息对应的最大回报值和所述获得最大回报值的车辆动作，设置所述
5 极限学习机网络模型的输入和输出；以及

网络训练模块，用于对所述极限学习机网络模型进行训练，获得所述极限学习机网络模型的输出权重。

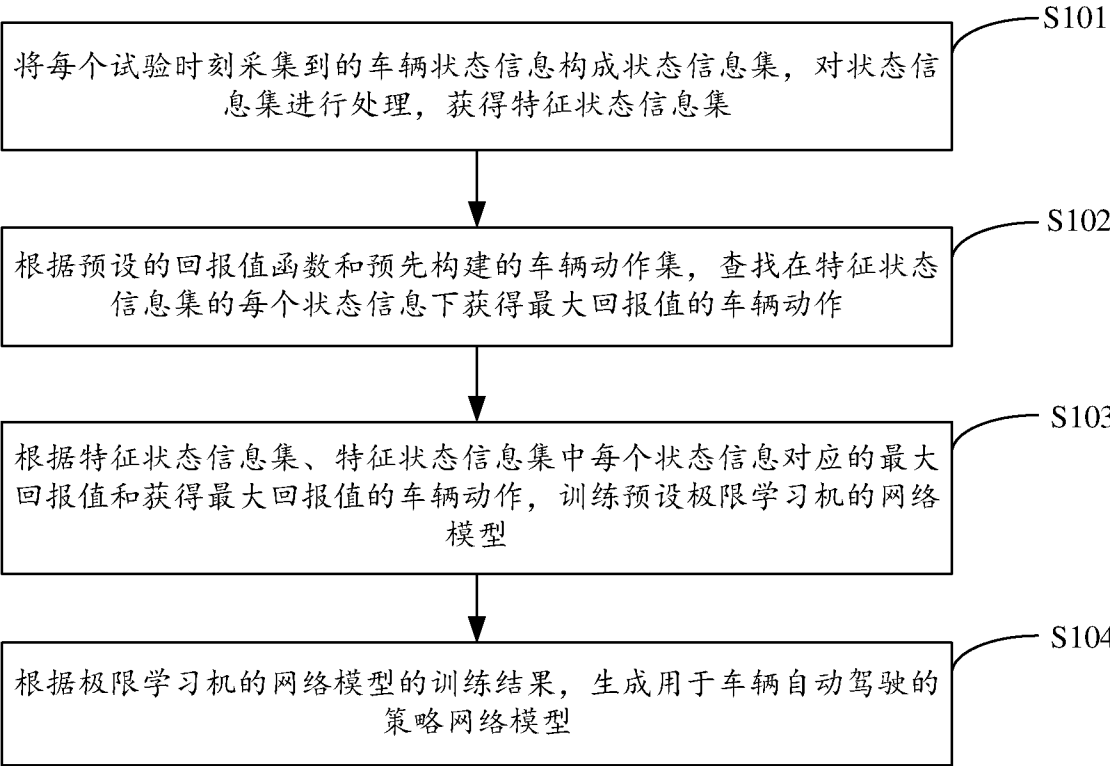


图 1

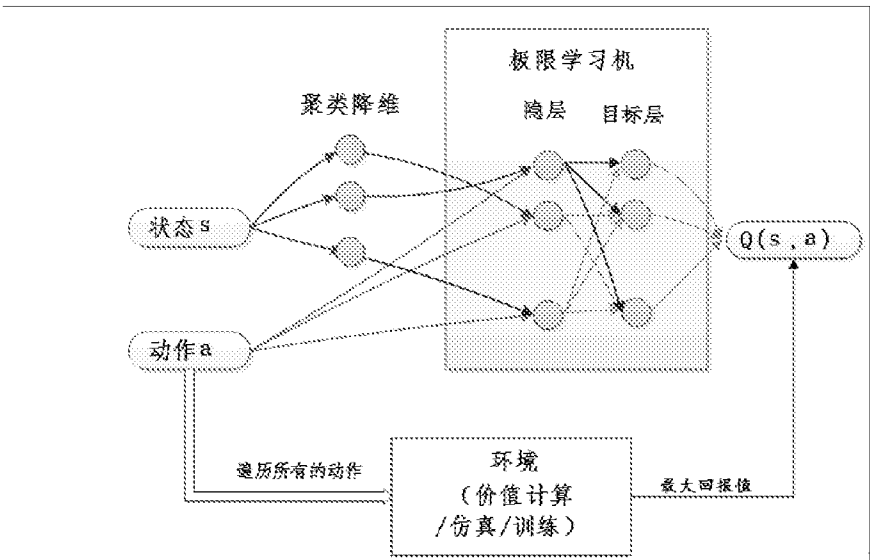


图 2

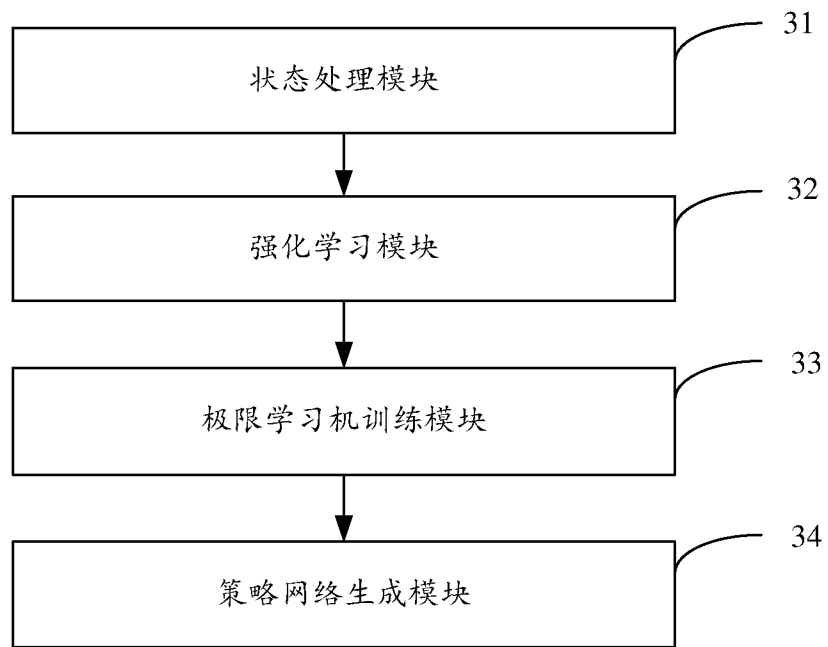


图 3

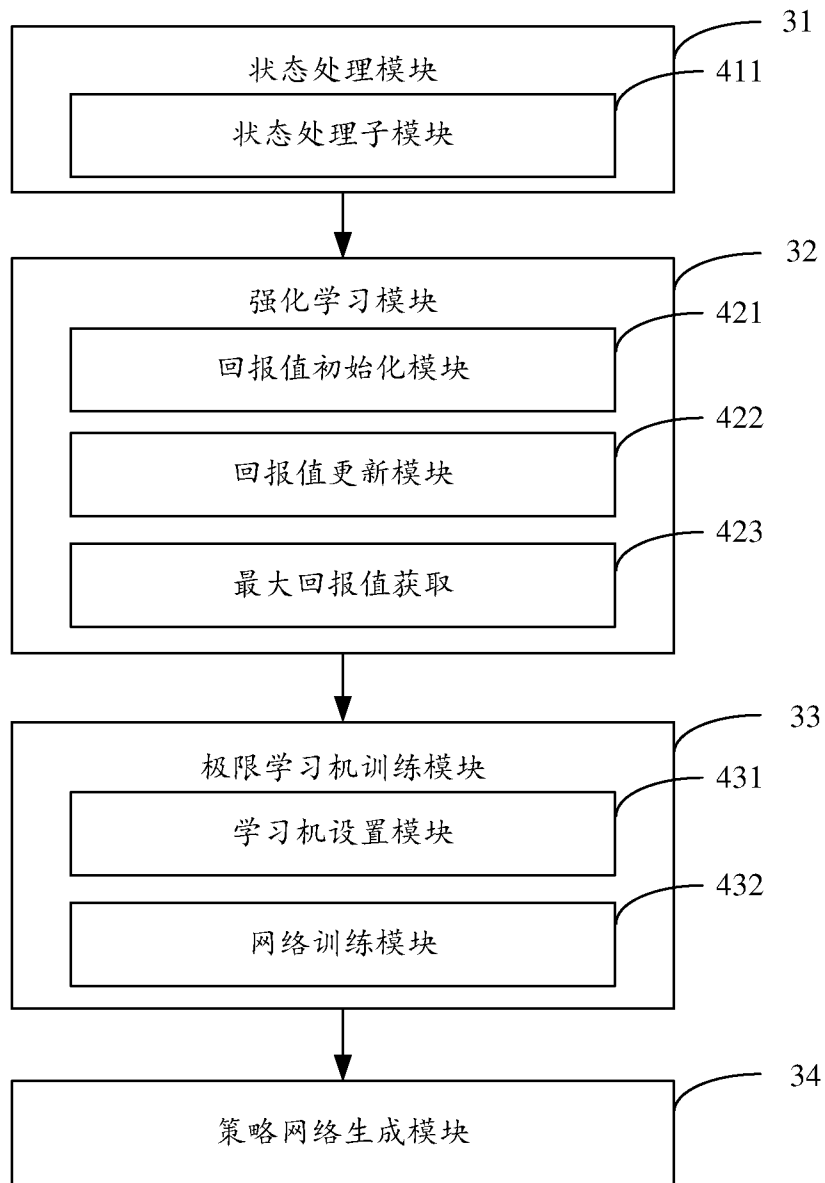


图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/084081

A. CLASSIFICATION OF SUBJECT MATTER

G06N 5/02 (2006.01) i; G06N 99/00 (2010.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNKI, VEN: 自动驾驶, 无人驾驶, 策略, 模型, 学习, 训练, automatic, driverless, drive, strategy, model, study, learn, educate, training

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E	CN 107229973 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 03 October 2017 (03.10.2017), claims 1-10	1-10
X	CN 106080590 A (BAIDU ONLINE NETWORK TECHNOLOGY BEIJING CO., LTD.) 09 November 2016 (09.11.2016), description, pages 3-7, and figures 1-4	1, 2, 5-7, 10
A	CN 106428000 A (TSINGHUA UNIVERSITY) 22 February 2017 (22.02.2017), entire document	1-10
A	CN 103605285 A (NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY) 26 February 2014 (26.02.2014), entire document	1-10
A	US 6493614 B1 (SAMSUNG ELECTRONICS CO., LTD.) 10 December 2002 (10.12.2002), entire document	1-10

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	
“A” document defining the general state of the art which is not considered to be of particular relevance	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“E” earlier application or patent but published on or after the international filing date	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	“&” document member of the same patent family

Date of the actual completion of the international search 17 November 2017	Date of mailing of the international search report 21 December 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer HAN, Wei Telephone No. (86-10) 62085295

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/084081

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 107229973 A	03 October 2017	None	
CN 106080590 A	09 November 2016	None	
CN 106428000 A	22 February 2017	None	
CN 103605285 A	26 February 2014	None	
US 6493614 B1	10 December 2002	KR 100504255 B1	28 July 2005
		KR 20030054256 A	02 July 2003

国际检索报告

国际申请号

PCT/CN2017/084081

A. 主题的分类 G06N 5/02 (2006.01) i; G06N 99/00 (2010.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类		
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06N 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNABS, CNKI, VEN 自动驾驶, 无人驾驶, 策略, 模型, 学习, 训练, automatic, driverless, drive, strategy, model, study, learn, educate, training		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
E	CN 107229973 A (中国科学院深圳先进技术研究院) 2017年 10月 3日 (2017 - 10 - 03) 权利要求1-10	1-10
X	CN 106080590 A (百度在线网络技术北京有限公司) 2016年 11月 9日 (2016 - 11 - 09) 说明书第3-7页以及图1-4	1, 2, 5-7, 10
A	CN 106428000 A (清华大学) 2017年 2月 22日 (2017 - 02 - 22) 全文	1-10
A	CN 103605285 A (南京理工大学) 2014年 2月 26日 (2014 - 02 - 26) 全文	1-10
A	US 6493614 B1 (SAMSUNG ELECTRONICS CO LTD) 2002年 12月 10日 (2002 - 12 - 10) 全文	1-10
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期		国际检索报告邮寄日期
2017年 11月 17日		2017年 12月 21日
ISA/CN的名称和邮寄地址		受权官员
中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10) 62019451		韩薇 电话号码 (86-10) 62085295

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/084081

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	107229973	A	2017年 10月 3日	无	
CN	106080590	A	2016年 11月 9日	无	
CN	106428000	A	2017年 2月 22日	无	
CN	103605285	A	2014年 2月 26日	无	
US	6493614	B1	2002年 12月 10日	KR 100504255 B1	2005年 7月 28日
				KR 20030054256 A	2003年 7月 2日