权 利 要 求 书

1、一种基于优先经验重放的深度强化学习频谱共享方法，其特征在于，所述方法包括如下步骤：

S1，构建频谱共享模型；

S2，在步骤 S1 中的频谱共享模型下，将频谱共享问题建模为深度强化学习中智能体与环境交互的马尔科夫决策过程，训练基于样本优先经验重放的深度强化学习模型，获得认知用户功率传输的学习价值信息；

S3，根据步骤 S2 中所获取的认知用户功率传输学习价值信息，判断频谱大数据下频谱共享控制决策，其中，所述控制决策实现了认知用户通过调节自身传输功率在不影响主用户通信质量下共享主用户的频谱。

2、根据权利要求 1 所述的基于优先体验重放的深度强化学习频谱共享方法，其特征在于，步骤 S1 中，频谱共享模型包括主用户和认知用户，它们以非协作的方式工作；主用户根据自身的功率控制策略更新发射功率，认知用户采用基于优先经验重放的深度强化学习机制更新发射功率来共享主用户的频谱；采用信干噪比 SINR 度量主用户和认知用户的服务质量 QoS；第 i 个接收机的 SINR_i 为：$\mathrm{SINR}_i=\dfrac{|h_{ii}|^2 P_i}{\sum_{j\neq i}|h_{ji}|^2 P_j+N_i}$，其中，$h_{ij}$ 表示发射端 i 到接收端 j 的信道增益，p_i 是第 i 个发射端的传输功率，N_i 表示第 i 个接收端噪声功率；假设主用户接收端和认知用户接收端成功接收传输功率必须满足一个最小 SINR，即：$\mathrm{SINR}_i\geq\mu_i, i=1,2$；主用户功率控制策略算法为：$P_{pu}(t+1)=\Im\left(\dfrac{\mu_1 P_{pu}(t)}{\mathrm{SINR}_1(t)}\right)$，其中，$P_{pu}(t)$ 表示在第 t 个时间帧主用户的传输功率，$\mathrm{SINR}_1(t)$ 表示在第 t 个时间帧主用户接收端测得的信干噪比，$\Im(\cdot)$ 表示一个离散化操作，目的是将一组连续的值映射到一组离散的值上，即 $p_{pu}(t)=\{p_{pu}^1,\cdots,p_{pu}^L\}$，其中 $p_{pu}^1\leq\cdots\leq p_{pu}^L$，且

$$\Im(x)=\begin{cases}P_{pu}^m & x>P_{pu}^m, m\in(1,L)\\ P_{pu}^L & x>P_{pu}^L\end{cases}。$$

3、根据权利要求 1 所述的基于优先经验重放的深度强化学习频谱共享方法，其特征在于，所述步骤 S2 中，所述基于优先经验重放的深度强化学习模型的训练过程如下：

S21，初始化经验池容量为 D，神经网络初始化；设定经验池为一个满二叉树，叶子节点可储存 D 个状态动作对；初始化 Q 网络的权重参数为 θ，目标网络 $\hat{Q}$ 的权重参数为 $\theta\_=\theta$；

S22，将频谱共享问题建模为深度强化学习中智能体与环境交互的马尔科夫决策过程，建立状态空间 $S(t)$，定义动作空间 A 以及即时奖赏 $r_{ss'}^a(t)$ 计算模型；

S23，积累具有优先级的经验池，具体步骤如下：

S231、初始化状态空间$S(1)$；根据当前输入状态$S(1)$，通过Q网络得到全部动作，利用ε贪心算法选取动作，具体是以ε的概率从动作空间A选择一个动作$a(t)$，否则以$1-\varepsilon$的概率选取最大Q值的动作$a_t=\max_{a_t} Q(s_t,a;\theta)$，其中$t$表示时间；

S232、根据步骤S1中主用户的功率更新策略更新主用户的传输功率，在执行动作$a(t)$后，得到即时奖励$r_{ss'}^{a}(t)$和t+1时刻的状态$S(t+1)$；

S233、将t+1时刻的状态$S(t+1)$作为当前输入状态，重复步骤S231和S232，将计算得到的状态动作对$e(t)\triangleq\{S(t),a(t),r(t),S(t+1)\}$和最大优先级$d_t=\max_{i<t} d_i$存到满二叉树构成的经验池中，满二叉树中只有叶子节点储存状态动作对；

S234、重复步骤S233直到经验池的D空间被储存满，经验池的满二叉树储存满后每执行一次步骤S233便跳转执行一次步骤S24；

S24，训练频谱共享模型下深度强化学习神经网络，具体步骤如下：

S241、从满二叉树中采样小批量O的$e(t)$，每个样本被采样的概率基于$j\sim D(j)=d_j^a\big/\sum_i d_i^a$，采样样本储存在一个$(m,n)$的二维矩阵，其中，$m$为样本容量大小，$n$为每个样本储存的信息数量，满足$n=2*s+a+1$，$s$为状态的维度，$a$为动作的维度，1为存储奖励信息的预留空间；

S242、对步骤S241中的小批量样本O计算每个$e(t)$采样样本$j\sim D(j)=d_j^a\big/\sum_i d_i^a$；

S243、对步骤S241中的小批量样本O计算每个$e(t)$样本重要性采样权重ω，采样权重主要是为了纠正网络训练过拟合问题，即：$w_j=(N\bullet D(j))^{-\beta}\big/\max_i w_i$，其中$\beta$表示纠正程度；

S244、计算步骤S241中所有样本的时序误差$\delta_j=R_j+\gamma\hat{Q}(S_j',\arg\max_{a'} Q(S_j',a;\theta);\theta_)-Q(S_j,A_j;\theta)$，并更新满二叉树中所有节点的优先级$d_j\leftarrow|\delta_j|$；

S245、使用均方差损失函数$L(\theta)=\frac{1}{o}\sum_{j=1}^{m} w_j(y_j-Q(S_j,A_j,\theta))^2$，通过神经网络的Adam梯度反向传播来更新$Q$网络的所有参数$\theta$；

S246、如果t是更新步长C的整数倍，更新目标网络$\hat{Q}$参数$\theta_=\theta$；

S247、如果$S(t+1)$是终止状态，当前训练完成，否则转到步骤S23。

4、根据权利要求1所述的基于优先经验重放的深度强化学习频谱共享方法，其特征在于，步骤3中包括，利用训练好的基于优先经验重放的深度强化学习模型在频谱大数据中实现频谱共享，具体包括以下几个步骤：

S31，初始化认知用户的传输功率，得到状态$S(1)$；

S32，选择动作 $a_t = \max_{a_t} Q(s_t, a; \theta^*)$ 得到 $S(t+1)$，即在 $t+1$ 时刻认知用户通过优先经验重放的深度强化学习模型智能更新传输功率，在不影响主用户的通信质量下共享主用户的频谱，其中 θ^* 为已训练神经网络的权重参数。

5、根据权利要求 3 所述的基于优先经验重放的深度强化学习频谱共享方法，其特征在于，所述步骤 S22 中，建立状态空间 S(t)具体过程如下：

选择传感器节点的接收功率作为状态空间，即：$S(t) = \left[p_1^s(t), \cdots, p_N^s(t) \right]^T$，其中 N 为频谱共享模型中传感器节点的数量；所述传感器节点用于辅助认知用户学习有效的功率控制策略而在频谱共享模型中而设置，所述传感器节点可用于测量在无线电环境中不同位置的接收信号强度，该信号强度由主用户和认知用户的传输功率控制，且只有认知用户可以访问；$P_n^s(t)$ 表示 t 时刻传感器节点 n 的接收功率，满足 $P_n^s(t) = P_{pu}(t) g_{pn} + P_{su}(t) g_{sn} + w_n(t)$，其中 $P_{pu}(t)$ 和 $P_{su}(t)$ 分别表示主用户和认知用户的传输功率，$w_n(t)$ 表示具有方差的零均值高斯随机变量，g_{pn} 和 (g_{sn}) 表示主用户(知用户)端与传感器节点 n 之间传输的路径损耗，满足 $g_{pn} = (\lambda / 4\pi d_{pn})^2$ ， $g_{sn} = (\lambda / 4\pi d_{sn})^2$，其中 λ 表示信号波长，$d_{pn}(d_{sn})$ 表示主用户(认知用户)发射端与传感器节点 n 的距离。

6、根据权利要求 3 所述的基于优先经验重放的深度强化学习频谱共享方法，其特征在于，所述步骤 S22 中，建立动作空间 A 过程如下：

选取认知用户的传输功率作为控制动作，即 $A(t) = P_{su}(t)$,其中 $P_{su}(t) = \left\{ P_{su}^1, \cdots, P_{su}^L \right\}$， $p_{su}^1 \le \cdots \le p_{su}^L$；认知用户通过在每个时刻 t 收集的传感器节点接收信号强度智能学习并调节自身传输功率，使得主用户和次用户能够在满足 QoS 需求下成功的传输数据。

7. 根据权利要求 3 所述的基于优先经验重放的深度强化学习频谱共享方法，其特征在于，所述步骤 S22 中，建立即时奖赏 $r_{ss'}^a(t)$ 计算模型的过程如下：

选取常数 C 作为即时奖励，当主用户接收端和认知用户接收端成功传输数据的同时都能够满足一个最小信干噪比要求时可获得奖励 C，即时奖励函数为：

$r_{ss'}^a(t) = \begin{cases} C & \mathrm{SINR}_1(t+1) \ge \mu_1 \cap \mathrm{SINR}_2(t+1) \ge \mu_2 \\ 0 & otherwise \end{cases}$，其中 $r_{ss'}^a(t)$ 指 t 时刻在状态 s 下采取动作 a 到状态 s' 的即时奖励。