

**《算法分析与设计》**

**课程实践报告**

**报告题目：基于多种智能方法的的斗地主AI算法**

**学院名称： 信息科学与技术学院**

**年 级： 2018级**

**授课教师：**  **杜圣东**

**辅导教师：**

|  |  |  |  |
| --- | --- | --- | --- |
| **学号** | **姓名** | **分工** | **成绩** |
| **2018112664** | **江胤佐** |  |  |
| **2018110610** | **吴玮** |  |  |
| **2018115053** | **孙硕人** |  |  |

**二〇二〇 年 五 月**

目录

[一． 选题动机 3](#_Toc43332496)

[1.1 为何选择机器学习 3](#_Toc43332497)

[1.2 为何选择斗地主 3](#_Toc43332498)

[二． 相关研究 3](#_Toc43332499)

[2.1 结合跳点引导的无人机随机搜索避撞决策方法 3](#_Toc43332501)

[2.2 命名实体识别算法 4](#_Toc43332502)

[2.3 CGF战术决策 4](#_Toc43332503)

[2.4 围棋人工智能AlphaGo 4](#_Toc43332504)

[三． 算法分析 4](#_Toc43332505)

[3.1 支持向量机 4](#_Toc43332507)

[3.2 蒙特卡洛树搜索 6](#_Toc43332508)

[3.3 强化学习 8](#_Toc43332509)

[四． 设计概述 8](#_Toc43332510)

[4.1 规则设计 8](#_Toc43332512)

[4.2 算法设计 8](#_Toc43332513)

[五． 实现步骤 8](#_Toc43332514)

[5.1 规则设计实现 8](#_Toc43332516)

[5.2 SVM二分类手牌叫地主算法实现 10](#_Toc43332517)

[5.3 蒙特卡洛树搜索树设计实现 11](#_Toc43332518)

[5.4 强化学习算法设计实现 12](#_Toc43332519)

[六． 结果分析 12](#_Toc43332520)

# 选题动机

## 为何选择机器学习

机器学习是研究怎样使用计算机模拟或实现人类学习活动的科学，是人工智能中最具智能特征，最前沿的研究领域之一，同样也是学习人工智能的基础与核心。通过对机器学习、人工智能相关方面的研究进行调研后，我们发现机器学习的一个重要算法——蒙特卡洛树搜索算法被广泛地应用在各种博弈对抗类游戏的AI上，从比较知名且复杂的AlphaGo到一些较为简单的五子棋，象棋等棋类都采用了该算法，训练出的AI甚至比人类更具博弈性和竞争力。

## 为何选择斗地主

斗地主是最少由3个玩家进行，用一副54张的扑克（包括王牌），其中一方为地主，其余两家为另一方，双方对战，先出完牌的一方获胜的卡牌游戏。之所以选择斗地主作为本次研究的对象，是因为与上文所提到的那些棋类博弈对抗游戏相比，斗地主是一种不完全信息博弈，由于是2对1的不公平对抗，而且双方手牌都是随机发放，这使得斗地主的对局更为不确定，这就需要牌手具有对当前牌局有深刻理解，对己方手牌，敌方手牌有明确的推断，所以对于传统的基于策略的算法而言，复杂度将会非常高，甚至难以实现，而这种高复杂度的场景正好适用蒙特卡洛算法这种近似推断方法。

我们小组希望通过对蒙特卡洛树搜索算法的调研，查阅资料，对其进行优化、调整，最终设计一种能够训练斗地主AI的算法，以更深入地了解机器学习算法的实际应用。

# 相关研究

## 基于蒙特卡洛树搜索的相关研究

1. **围棋人工智能AlphaGo**

阿尔法围棋（AlphaGo）是第一个击败人类职业围棋选手、第一个战胜围棋世界冠军的人工智能机器人。阿尔法围棋用到了很多新技术，如神经网络、深度学习、蒙特卡洛树搜索法等。阿尔法围棋系统主要由几个部分组成：策略网络（Policy Network），给定当前局面，预测并采样下一步的走棋；快速走子（Fast rollout），目标和策略网络一样，但在适当牺牲走棋质量的条件下，速度要比策略网络快1000倍；价值网络（Value Network），给定当前局面，估计是白胜概率大还是黑胜概率大；蒙特卡洛树搜索（Monte Carlo Tree Search），把以上这四个部分连起来，形成一个完整的系统。

1. **CGF战术决策**

计算机生成兵力(Computer Generated Forces, CGF)的行为建模一直是作战仿真技术中的重点和难点。在一些典型的CGF系统开发实践中,行为模型需要大量地依赖领域相关人员参与构建,以形成足够的领域知识驱动复杂行为。但是这些模型主要产生反应式行为,除了构建过程繁琐之外,也无法对人的高级认知活动(如态势感知、任务规划和推理决策等)进行有效表示。将蒙特卡洛树搜索应用于CGF战术决策能够使CGF在虚拟战场中表现出更具智能性的行为。同时,其在推理过程中考虑了敌我双方的可能行动,在此基础上形成对不同方案的评估,这种方式非常适合分析类似作战这种存在动态对抗的环境。

## 支持向量机的相关背景

1. **文本信息监测**

基于集成学习策略SVM分类法对中文文本中具有欺骗性的信息进行检测识别. 通过改进的二分k-均值划分法划分实验中的训练样本, 进而对各个训练样本子集构建相应的分类器, 称之为子分类器. 最终通过集成所有子SVM分类器的结果确定实验的分类效果, 研究人员对其良好的识别率进行分析总结。

1. **手写文档分类**

通过结构化学习技术确定SVM参数, 并构造松弛结构SVM训练估计最优的参数. 结构化SVM方法降低识别的计算复杂度并解决了手写文档不规则和多样化的干扰因素. 该文仅在拉丁语和印度语文档中证明其可靠性, 可通过将该方法应用于英文以及汉语文本中, 扩大其文本应用范围。

## QLearning的相关背景

由DeepMind公司开发的利用深度卷积神经网络来进行Q-Learning的算法。在使用非线性函数逼近的时候，强化学习经常会有不稳定性或者发散性：这种不稳定性来于当前的观测中有比较强的自相关。DeepMind 通过使用经历回放，也就是每次学习的时候并不直接从最近的经历中学习，而是从之前的经历中随机采样来进行训练。



# 算法分析

## 支持向量机

支持向量机（support vector machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM学习算法就是求解凸二次规划的最优化算法。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。

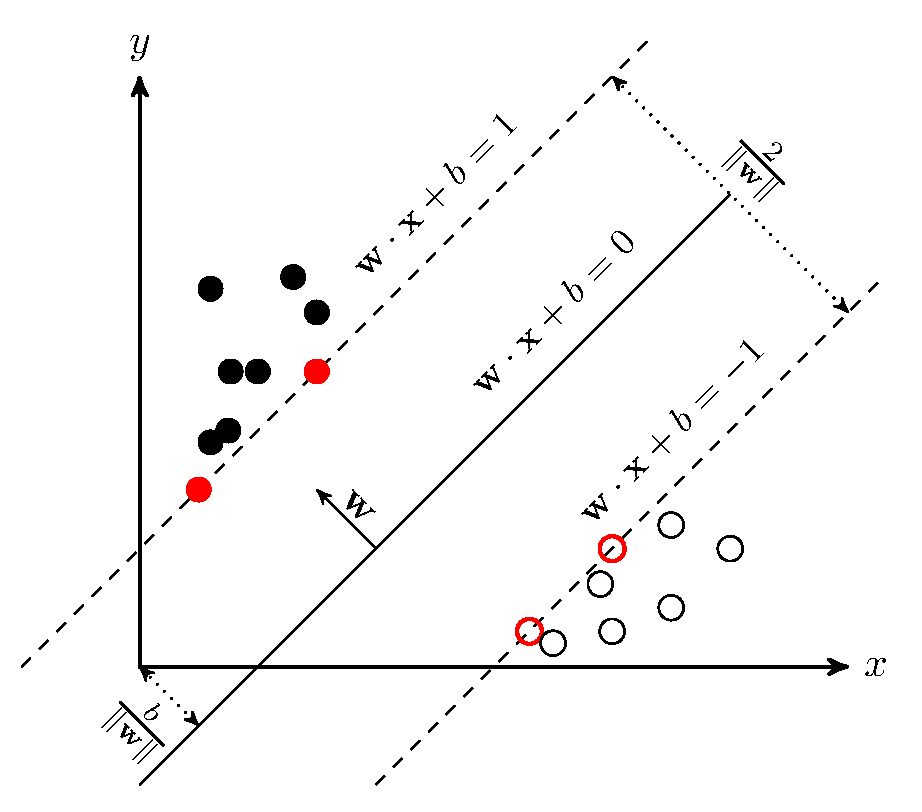


图3.1（1） 几何间隔最大的分离超平面

具体的算法原理说明由于过于复杂在此按下不表。

直接给出线性支持向量机学习算法：

**输入：**训练数据集其中，

**输出：**分离超平面和分类决策函数

1. 选择惩罚参数

选择惩罚参数，构造并求解凸二次规划问题

得到最优解

1. 计算

选择的一个分量满足条件,计算

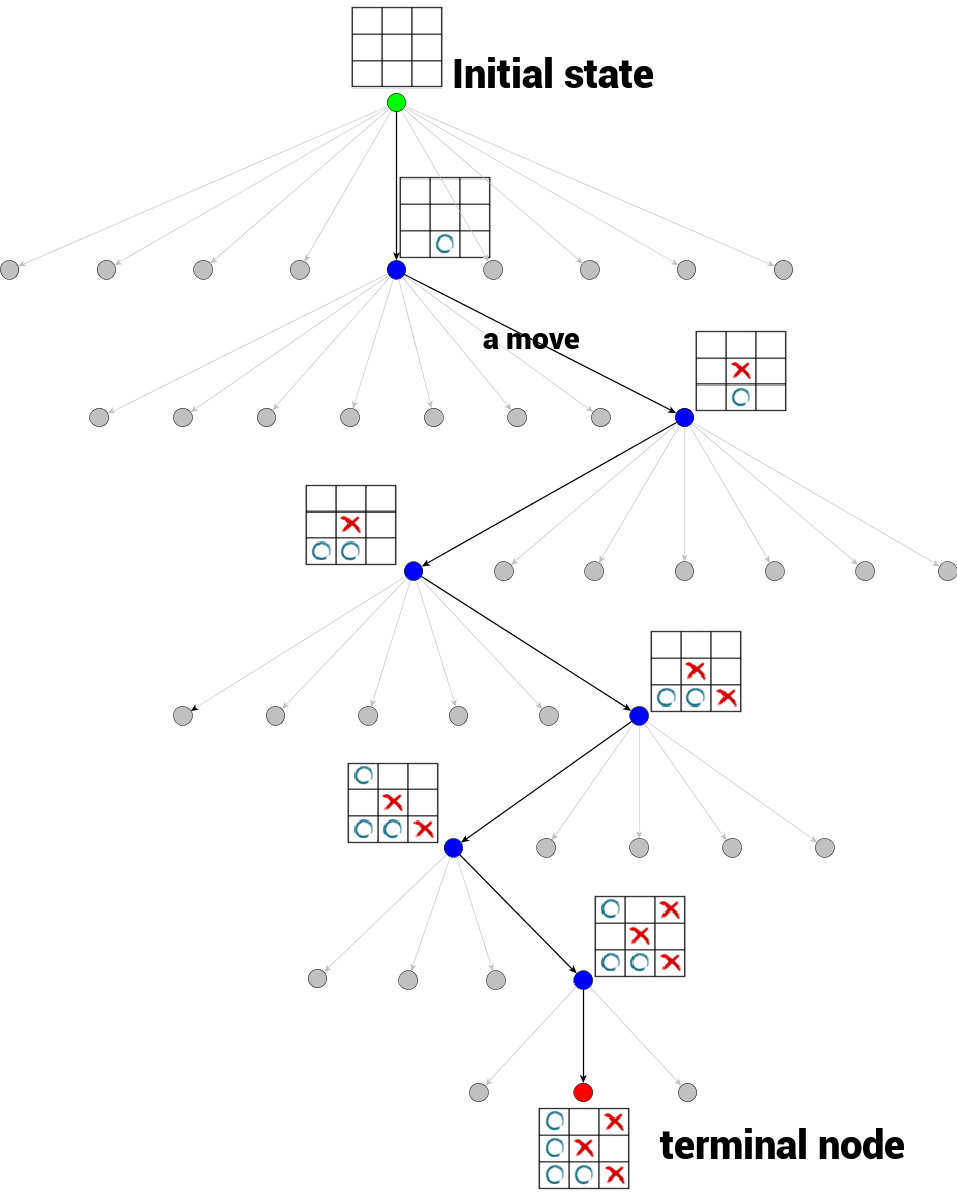
1. 求分离超平面

分类决策函数：

## 蒙特卡洛树搜索

如图3.2(1)我们采用博弈树(Game Tree)来表示一局游戏，定义如下：

1. 每个结点(Node)都代表一个状态(State)
2. 从一个结点移动一步，将会到达它的子节点(Children Node)
3. 子节点的个数叫做分支因子(Branching Factor)
4. 根节点(Root Node)表示初始状态(Initial State)
5. 终止节点(terminal nodes)则没有子节点。



3.2(1) 一局游戏的表示

每次都是从初始状态、树的根结点开始。在tic-tac-toe游戏里面初始状态就是一张空的棋盘。从一个节点转移到另一个节点叫作一个move。分支因子(Branching Factor)，tic-tac-toe中树越深，分支因子也越少，也就是子节点(Children Node)的数量越少。游戏结束表示终止节点。从根节点到终止节点一次表示一局单个游戏。

我们希望找到的就是最佳策略(The Most Promising Next Move)。如果你知道对手的策略那你可以针对这个策略求解，但是大多数情况下是不知道对手的策略的，所以我们需要用Minimax的方法，假设你的对手是非常机智的，每次他都会采取最佳策略。

假设A与B博弈，A期望最大化自己的收益，因为是零和博弈，所以B期望A的收益最小，Minimax算法可描述为如下形式：

和是玩家A和B的效益函数。

表示从当前状态和采取的动作转移到下一个状态。

评估最终的游戏分数。

是最终的游戏状态。

简单地说，就是给定一个状态期望找到一个动作在对手最小化你的奖励的同时找到一个最大化自己的奖励。

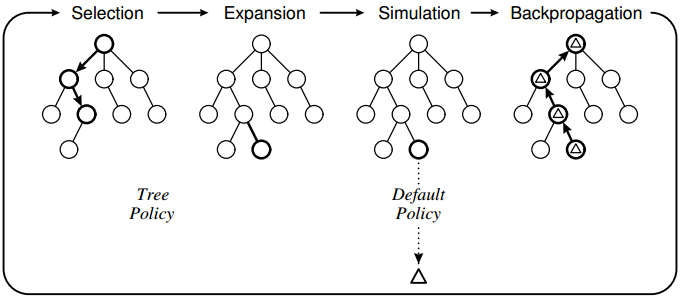
Minimax算法最大的弱点就是需要扩展整棵树，对于高分支因子的游戏，像围棋、象棋这种，算法就很难处理。

对于上述问题的一种解决方法就是扩展树结构到一定的阈值深度(Threshold Depth)。因此我们需要一个评估函数，评估每一个非终止节点。

另一种解决树扩展太大的方法就是Alpha-Beta剪枝算法。它会避免一些分支的展开，它最好的结果就是与Minimax算法效果相同，因为它减少了搜索空间。

蒙特卡洛通过多次模拟仿真，预测出最佳策略。最核心的东西就是搜索。搜索是对整棵博弈树的组合遍历，单次的遍历是从根结点开始，到一个未完全展开的节点(Not Full Expanded)。未完全展开的意思就是它至少有一个孩子节点未被访问，或者称作未被探索过。当遇到未被完全展开过的节点，选择它的一个未被访问的子节点作为根结点，进行一次模拟。仿真的结果反向传播(Propagated Back)用于更新当前树的根结点,并更新博弈树节点的统计信息。当整棵博弈树搜索结束后，就相当于拿到了这颗博弈树的策略。

基本的 MCTS 算法非常简单：根据模拟的输出结果，按照节点构造搜索树。其过程可以分为下面的若干步：



1. 选择(Selection)：从根节点 R 开始，递归选择最优的子节点（后面会解释）直到达到叶子节点 L。
2. 扩展(Expansion)：如果 L 不是一个终止节点（也就是，不会导致博弈游戏终止）那么就创建一个或者更多的字子节点，选择其中一个 C。
3. 模拟(Simulation)：从 C 开始运行一个模拟的输出，直到博弈游戏结束。
4. 反向传播(Backpropagation)：用模拟的结果输出更新当前行动序列。

## 强化学习

未完待续

# 设计概述



## 规则设计

想要让计算机理解什么是斗地主，我们首先要为其构建斗地主的规则，包括：

1. 牌型规则：明确所有的牌型组合
2. 出牌规则：明确不同牌型间的大小比较与炸弹王炸等特殊换手机制
3. 角色规则：明确地主与农民两个角色的差异
4. 胜利规则：明确胜利的条件为出光手牌

## 算法设计

斗地主的游戏过程主要由三个主要过程：出牌、跟牌、其他（包括规则定义、叫地主、出牌大小判断、出牌合法性判断、发牌等等）。

由于使用手牌作为状态，可能性过多，导致算法可行性较差，所有需要另外进行动作与状态的定义。

叫地主部分使用支持向量机学习实现。

牌型（出牌合法性）判断与出牌大小判断由自定义算法实现（具体设计参见5.1 规则设计实现）

在出牌跟牌部分（即涉及到博弈的部分）需要使用蒙特卡洛树搜索算法。

在模型训练时，需要使用强化学习算法使AI自我对战。

# 实现步骤



## 规则设计实现

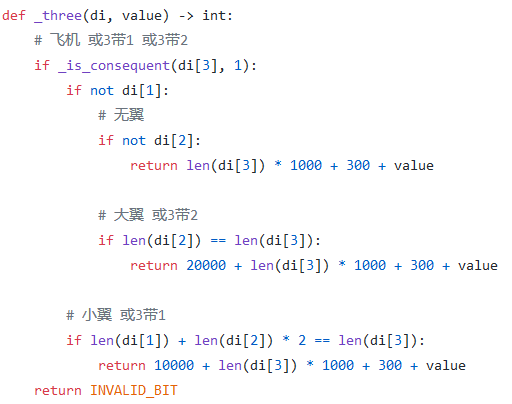
1. **牌型规则**

规则定义：根据斗地主规则定义了各种牌型的大小（每种牌型对应一个不同大小的值，对于同一牌型，值越大表示该牌越大）。

1. **出牌规则**

出牌大小判断：根据牌型规则集不同牌型所代表值对当前出牌及上家出牌大小进行比较，如果大于则合法。

出牌合法性判断：判断所出手牌是否为提前定义的牌型规则集的子集。比如：单牌顺子一次性不能少于5张且不能带2，出“3456”、“JQKA2”都会判定为不合法。



5.1(1)牌型定义、出牌大小及合法性判断代码示例

拆牌：拆牌是在不考虑其他玩家手牌的情况下对当前可打出的所有牌型进行拆分，状态(state)表示目前的手牌，动作(action)表示待出的牌，Q值表示状态-动作的价值，Q(s, a)值越大，则越要在状态s下采取行动a。

拆牌具体实现步骤：

(1)对每一种状态-动作计算其d值：对于所有动作，获取状态做出动作后的下一步状态，即去除出掉当前牌后剩余牌组的状态，若该动作做完不空，则加入下一状态的d值，直到该动作做完后牌组为空，则得到最大的d值。

（计算d值：首先获取当前初始牌的d（分解）值（先不考虑几带几的牌），获取当前牌中小于2的牌后，得到不同张数的牌有哪些，比如说在这副牌中单张的牌有3、4、5，而两张的牌有66、77等。再对不同张数的牌类型，将不连续的牌分解出来，得到d值。）

(2)获取较好的动作：将手牌分解成不连续的部分（同上一步分解不连续的牌的操作），之后获取小于2的牌中当前状态下较好的拆牌，得到最终拆牌结果。

（获取小于2的牌中当前状态下较好的拆牌：对获取所有单种牌面的动作（单，对，三，炸弹）以及获取顺子/连对/飞机/炸弹的动作（单，对，三，炸弹）进行评估，得到最优结果。

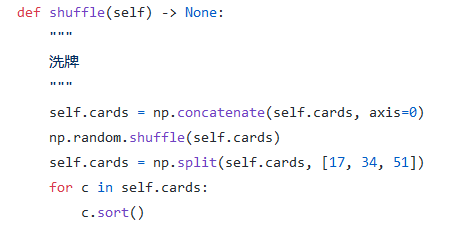
对每一个动作进行评估：对每一个动作对应的下一个状态根据之前列出的公式计算Q值，选取Q值最大的动作序列。）

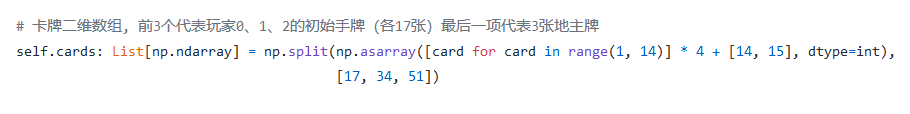
出牌、跟牌：通过蒙特卡洛树搜索对当前手牌拆牌所得的所有可能性牌型进行分析判断，选取对于当前局势获胜概率最大的一种牌型进行出牌或者跟牌，如果没有可以大过上家或胜率较低的手牌则选择不出；且玩家出牌后，将从当前手牌中删除之前出过的牌型。

（出牌、跟牌具体内容参见5.2蒙特卡洛树搜索树设计实现）

1. **角色规则**

发牌：首先将一副牌随机打乱，之后针对不同角色进行发牌，采用在一整副扑克牌中对三个玩家每个玩家随机发17张牌的方式，当有任意玩家叫地主时，则将剩下的三张牌发给叫地主的玩家。

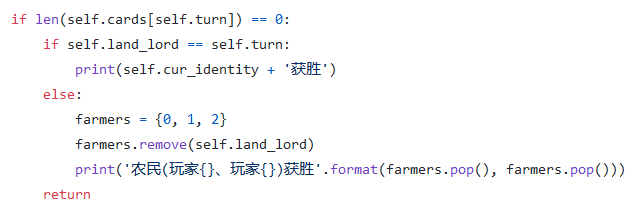




5.1(3)发牌代码示例

1. **胜利规则**

判断胜利：当有任何一人的手牌列表为空即判断其所属阵容（地主或农民）获胜。



5.1(4)判断胜利代码示例

## 动作与状态定义

1. **出牌动作**

AI出牌时的动作被化简为表5.2(1)：（其中的大小是针对符合规则的牌型值的序列而言的）

表5.2(1) 出牌动作定义表

|  |  |
| --- | --- |
| **值** | **动作** |
| X\*1（x取[0,4]） | 出最小、小、中、大、最大的单牌 |
| X\*2（x取[1,3]） | 出小、中、大的对子 |
| X\*3（x取[1,2]） | 出最小、最大的三张相同的牌 |
| X\*4（x取[1,3]） | 出小炸弹或大炸弹或王炸 |
| X\*5（x取[1,2]） | 出较小的或较大的顺子 |
| 6 | 出其它连对顺子飞机 |
| 7 | 出4带2 |

1. **出牌状态**

将AI当前出牌状态转换为一个长度为12的特征向量，如表5.2(2)：

表5.2(2) 当前出牌状态定义表

|  |  |  |
| --- | --- | --- |
| **状态说明** | **属性名** | **取值范围** |
| f1\_min(单张) | solo\_min | [0, 3] |
| f1\_max(单张) | solo\_max | [0, 3] |
| f2\_min(对子) | pair\_min | [0, 2] |
| f2\_max(对子) | pair\_max | [0, 2] |
| 三的数量(大于2记作2,含飞机) | trios | [0, 2] |
| 最大单顺（长为5） | seq\_solo\_5 | [0, 2] |
| 存在其它牌型(不含4带2) | other\_seq\_count | [0, 1] |
| 炸弹数量（大于2记作2） | bomb\_count | [0, 2] |
| 是否有王炸 | rocket | [0, 1] |
| 玩家位置 | player | [0, 2] |
| 上家手牌数-1 | hand\_p(大于5都记作5) | [0, 5] |
| 下家手牌数-1 | hand\_n(大于5都记作5) | [0, 5] |

1. **跟牌动作**

AI跟牌时的动作被化简为表5.2(3)：

表5.2(1) 跟牌动作定义表

|  |  |  |
| --- | --- | --- |
| **状态说明** | **属性名** | **取值范围（均为整数）** |
| 最佳拆牌的delta\_q | delta\_q | [0, 5] |
| 玩家身份 | player | [0, 2] |
| 上一个出牌者 | last\_combo\_owner | [0, 2] |
| 上家手牌数-1 | hand\_p(大于5都记作5) | [0, 5] |
| 下家手牌数-1 | hand\_n(大于5都记作5) | [0, 5] |

1. **跟牌状态**

将AI当前跟牌状态定义为表5.2(4)：

表5.2(4) 当前跟牌状态定义表

|  |  |
| --- | --- |
| **值** | **动作** |
| 0 | 空过 |
| 1-4 | 跟最小、较小、较大、最大 |
| 5 | 可能拆坏牌的情况下跟最大 |
| 6 | 小炸弹 |
| 7 | 大炸弹 |
| 8 | 王炸 |

## SVM二分类手牌叫地主算法实现

通过研究实际的斗地主高端对局发现，牌手叫地主的意向与当前手牌中的大小王（王炸）情况，除大小王以外的炸弹数量，手牌2的数量有着直接的关系。

1. **数据采集与处理**

通过采集斗地主顶级牌手的战报，记录超过四千组手牌与叫地主策略，完成数据集转化后，将数据集切分为训练集与测试集

1. **训练SVM分类器**

采用sklearn算法包，选取linear线性核进行训练

1. **计算SVC分类器的准确率**

采用测试集进行验证，计算出分类器的准确率为：0.932，符合预期

1. **绘制图像**

在坐标范围内以x，y，z轴分别表示三个特征，绘制散点图

## 蒙特卡洛树搜索树设计实现

如图5.3(1)所示，对于本算法而言一局斗地主中的一次出牌的决策流程大致可以分为如下步骤：

1. 更新当前游戏信息
2. 牌型拆分
3. 可选牌型枚举
4. 蒙特卡洛树搜索算法
5. 最佳出牌

如此循环直至游戏结束。其中蒙特卡洛树搜索算法便是其中的核心，利用它才能给出每次的最佳决策，直至获得最终胜利

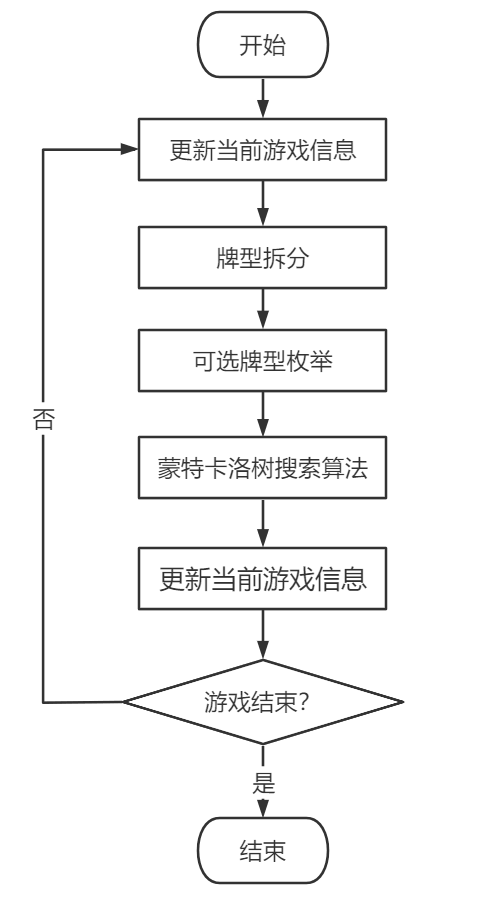


图5.3(1) 出牌的决策流程

很显然，在“斗地主”游戏中，其他玩家手牌信息对于当前玩家而言是隐藏的，这无疑会大大增加了博弈树中的节点数，导致博弈树空间进一步增大，增加解决“斗地主”游戏的难度.。所以普通的博弈树搜索算法，对于“斗地主”这类不完全信息博弈游戏不能提供很好的解决方法。针对问题规模大的问题，在斗地主游戏中，玩家根据自己手牌、两个玩家已出扑克以及两个玩家手牌张数等信息，通过对游戏未完成的博弈过程进行不断的模拟，以确定在当前情况下玩家最佳的决策(出牌)。

具体决策流程如下:

步骤１首先，根据当前玩家的手牌，以及所有玩家已出扑克等信息，基于“斗地主”的较小拆分算法，求解出待出牌集合，再将求解结果中的每个元素增加到博弈树中作为叶子节点.。

步骤２选择过程：在博弈树中所有的叶节点中，选取一个叶节点作为本次模拟的根节点。在选取的时候，由于需要权衡探索和利用的关系，所以将博弈树中的叶节点通过UCT算法计算该节点的选择评估值γ，选择评估值γ最大的叶节点作为新一次模拟的根节点进行模拟。UCT算法为：

式中表示节点i的选择评估值，表示节点i的平均收益，是常数,其作用是为了平衡探索和利用。是以第i个节点作为模拟搜索的根节点的次数。

步骤３扩展过程：在“斗地主”的过程中，由于“斗地主”游戏属于不完全信息博弈的一种，在对其叶节点进行扩展的时候，选中的叶节点的子节点数量较大。因此，在本游戏中，扩展阶段只将本次抽样模拟的根节点加入博弈树中。

步骤４模拟过程：从抽样的博弈环境信息开始，３个玩家根据手牌在满足游戏规则的前提下，随机进行交替出牌，直到到达博弈终止状态。

步骤５反馈过程：游戏到达终止状态后，将收益值返回到根节点，并沿途更新节点的统计估计值。实验过程中，在进行蒙特卡洛抽样过程中涉及到随机性问题，但其结果收敛性与实验环境无关，因为根据大数定律可知：当抽样次数足够多时，抽样收益的均值趋近于期望值。在“斗地主”游戏中的决策主体身份分为地主和农民，但博弈过程中，无论智能体身份为地主或农民，进行蒙特卡洛抽样的时候，对于其身份仅仅在最终反馈收益时加以区分，在抽样的过程中，所有决策主体无任何区别.

## 强化学习算法设计实现

未完待续

# 结果分析

未完待续