有无人协同互信置信度评估算法模型开发项目实施报告

# 项目名称

有无人协同互信置信度评估算法模型开发

# 项目需求/技术要求响应

有无人协同互信置信度评估算法模型开发项目开展智能化决策互信机制与分布式决策机制运行机理的研究，构建面向未来战场的有人无人协同作战分布式决策互信机制，搭建分布式协同互信决策模型，并面向分布式决策完成数字仿真验证。

## 研究内容

为构建有无人协同作战分布式决策互信机制，开展智能化决策互信机制与分布式决策机制运行机理的研究，开展智能化决策互信算法的开发与分布式协同互信决策模型搭建，对有无人协同分布式决策互信机制与运行机理进行数字仿真验证，具体完成的内容包括:

a) 多智能体互信协作的结构化建模以及相对应多智能体训练方法研究;

b) 多智能体协同方式及冲突消解方法研究;

c) 多智能体间互信机制及置信度计算方法研究;

d) 有无人协同互信置信度评估算法模型的开发与测试;

e) 面向分布式决策的模型集成及全流程仿真验证。

## 功能要求

a) 能够给出置信度评估结果;

b) 能够计算置信度参数;

c) 能够在预测决策冲突的情况下消解冲突;

d) 能够给出多智能体的目标分配方案、机动行为方案等;

e) 能够对分布式系统决策过程进行仿真验证

f) 能够给出仿真验证结果。

## 2. 性能要求

a) 算法模型中可容纳5类20个节点同时进行计算；

b) 算法模型计算出的置信度参数>85%。

c) 数据输入模型至模型输出结果的计算时间＜1 min。

# 技术要求分解

## 多智能体互信协作的结构化建模以及相对应多智能体训练方法研究

本项目面向未来构想的典型的多对多有无人机协同的空战场景(红方有人机包括2队共4架，无人机包括2架预警机、8架多用途战斗机、6架灵巧无人机；蓝方包括2个作战部队，分别为4架PCA、1架B-21，共8架PCA，2架B-21)，采用空战规则、深度学习、强化学习、专家模型等多种建模方法，生成不同飞机编队间的目标分配、机动性为决策等控制机制，为有无人协同智能决策分布式系统构造竞争对手，并优化智能分布式策略，实现模型自动化作战。同时通过计算有无人协同分布式系统各节点之间的决策互信度，在高置信度的前提下按需进行决策中的冲突消解，从而提前避免决策冲突的出现，有效提高有无人协同系统的运行流畅度。

针对上述场景，本项目模型通过采用强化学习中Actor-Critic模型框架，使用深度强化学习的方法，结合层次认知理论的思想，并辅以领域专家所提供的成熟规则及经验知识，构造智能分布式决策训练系统。具体的研究内容如下：

1. 基于深度强化学习的大规模多智能体强化学习决策模型的建立与训练方法研究，该模型支持动态组合方式的智能体协同决策；
2. 多智能体深度强化学习的奖励机制以及协同环境下共同获取奖励值得分配方法研究，解决多智能协同决策中的奖励分配问题；
3. 混合任务、博弈场景下的大规模智能体均衡策略优化方法研究，通过理论与实验方法证明算法可以有效地逼近纳什均衡解。使智能体在策略提升过程中具备较强的决策鲁棒性，避免拟合至某一特定策略与陷入局部最优；
4. 去中心化的多智能体组织形式下，在局部可观测（POMDP，DEC-POMDP）情况下，基于循环神经网络（RNN，LSTM）解决状态的不完全可观测性，以及在状态空间、样本量较大情况下对友方策略方法的拟合方法学习。
5. 多智能体分层结构下，离散-连续混合状态与动作空间下，大规模智能体强化学习神经网络模型建立、网络梯度连接与相应样本存储与训练方法研究；
6. 大规模智能体安全强化学习（safe-RL）方法研究；
7. 基于专家经验的模仿学习（GAIL）方法研究；
8. 基于好奇心、内部奖励机制（Intrinsic reward）等方法的多智能体对环境的探索方法研究；
9. 针对大规模空中对抗智能体分组协作对抗想定，进行大规模智能体异步决策方法模型建立、样本序列存储与智能体优化方法研究。
10. 对于认知层次理论（cognitive hierarchy theory）的研究，从多智能体的初级策略中抽取出多阶认知，以最终做出决策。

### MADDPG框架的研究

MADDPG算法在解决多智能体问题时具有以下三点特征：

1. 通过学习得到的最优策略，在应用时只利用局部信息就能给出最优动作。

2.不需要知道环境的动力学模型以及特殊的通信需求。

3.该算法不仅能用于合作环境，也能用于竞争环境。

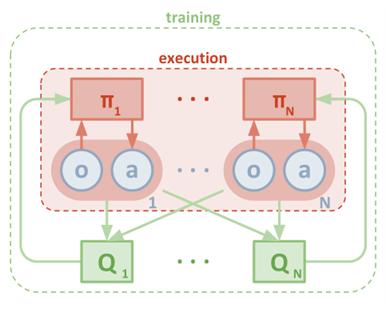


图2.2 MADDPG算法模型

MADDPG算法架构中每一个agent使用自己独立的Actor，通过自己观测状态o，输出确定的动作a，同时训练数据也只使用自己产生的训练数据，每一个agent同时也对应一个Critic，但是该Critic同时接收所有Actor产生的数据，像这样的Critic也是每个agent都有自己独立的一个，算法结构如图2.2所示。

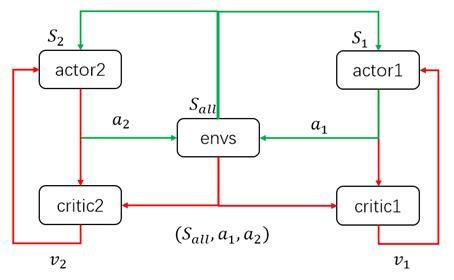


图2.2 Critic-Actor算法示例

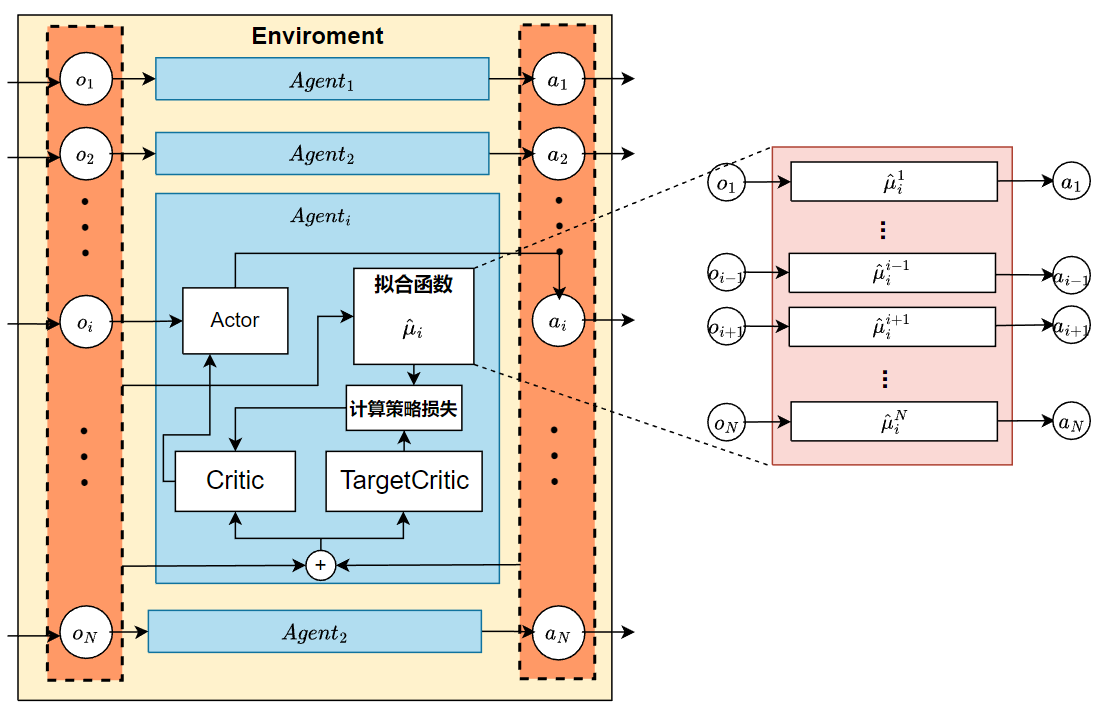
上面这张图以两个agent为例, 当模型训练好后，只需要两个actor与环境交互，即只需要绿色的循环。这里区别于单个agent的情况，每个agent的输入状态是不一样的。环境输出下一个全信息状态 后，actor1和actor2只能获取自己能够观测到的部分状态信息 ，​。

而在训练过程中，critic1和critic2可以获得全信息状态，同时还能获得两个agent采取的策略动作，。也就是说，Actor虽然不能看到全部信息，也不知道其他Actor的策略，但是每个Actor有一个上帝视角的导师，这个导师可以观测到所有信息，并指导对应的Actor优化策略。

MADDPG算法维护了四个部分的全连接神经网络，各部分神经网络的输入输出信息如下表所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 算法结构 | 神经网络 | 公式表示 | 输入 | 输出 |
| Critic | Target Critic Network |  | 1. 下一时刻系统的状态描述 2. 下一时刻智能体1到N的动作 | 输出用于计算target值的 （类似于DQN算法）  其中target值的计算公式为： |
| Critic Network |  | 1. 当前时刻系统的状态描述 2. 当前时刻智能体1到N的动作 | 输出为动作-状态值函数Q，分别用于两个部分：   1. 首先用于计算TD error，更新Critic 2. 用于Actor部分的参数更新 |
| Actor | Target Actor Network |  | 下一时刻自身智能体观察的状态 | 输出为 ，用于target Critic Network部分计算 |
| Actor Network |  | 当前时刻自身智能体观察的状态 | 输出为 ，和Critic Network输出的Q一起更新Actor神经网络的参数 |

而MADDPG框架的结构如图所示：

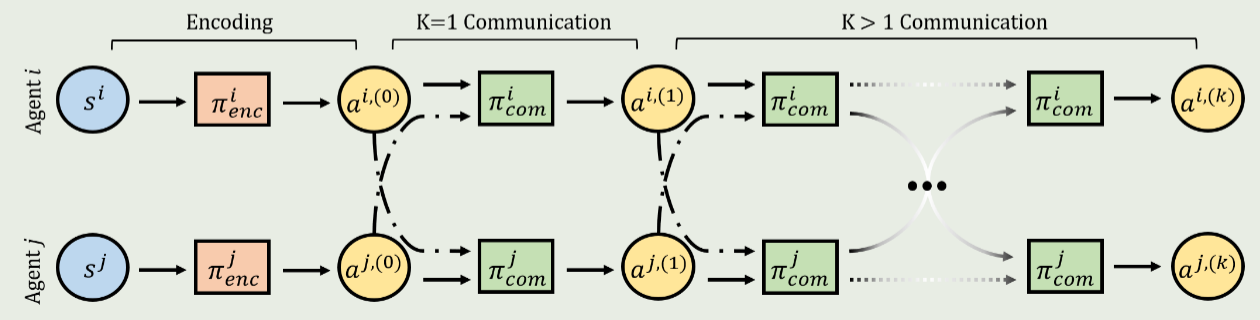


算法的Actor部分使用的是类似于DPG算法的更新方法，训练一个能够通过智能体自身观察及时做出策略动作的动作策略函数 。

而Critic部分使用 Q-learning的训练方法，学习一个评价函数，该评价函数在运行过程中需要其他智能体策略动作的拟合函数的参与：

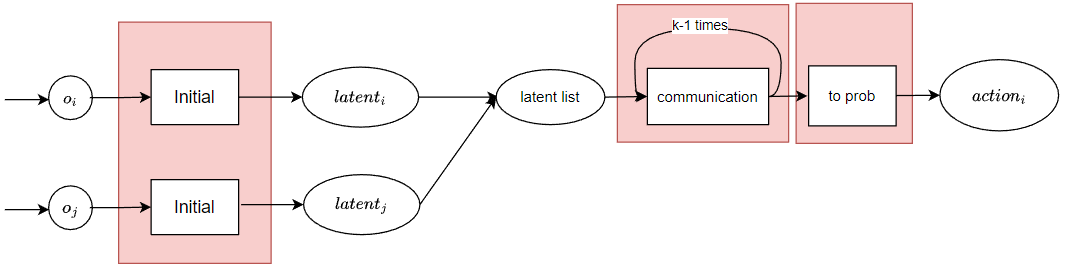
其中 是具有延迟参数 的目标策略集合。

### 基于InfoPG的互信息提升方法



利用认知层次理论和战略博弈论，其中每个主体对其邻居在有限理性下可能采取的行动具有k级策略。k级推理假设战略博弈中的agent不是完全理性的，因此每个agent的决策都通过合理化的层次结构迭代对其他 agent可能行动的预测。在 InfoPG 中，我们让agent能够在 k 个迭代推理步骤中与其潜在的猜测动作分布进行通信，并在 k 级合理化他们的动作决策，以最好地响应他们的队友级别 k-1 动作。我们通过上标 (k) 表示代理的合理性水平，其中 。将 表示为agent j 的 level-k 策略。可以证明在 k 级推理下，agent i 在 k + 1 级的策略 正是代理 i 对代理 j 的策略 的最佳响应。理论上，这个过程可以迭代地进行，直到我们得到 ，这对应于达到平衡策略。但实际运行中需要的运算量太大，大多情况下只会取到k=1或k=2。

根据InfoPG的思想，我们将模型的Actor部分修改成如下结构：



Actor部分由上图红色的三个部分组成，以下分别表述为initial层，communication层以及to prob层。其中：

* Initial层和原始MADDPG的Actor是一致的，输入为 的观察信息 ，输出初始动作 。如果直接将 用来执行并参与critic 函数，就是MADDPG的模型架构。
* Communication层将邻居Agent的 latent 拟合计算出来，经过 GRU网络，计算出自己在队友进行k-1级别动作时的最佳动作策略。
* To prob层的作用是将communicated的结果规范化整合为action输出格式。

这样一来，Actor部分的优化目标变为：

其中，为Critic批评函数得到的状态Q值。该方法被称为InfoPG的蒙特卡洛PG方法变体。该优化目标可以有效提升智能体决策之间的互信息下限。互信息（Mutual Information）是信息论中的一个概念，指的是随机变量之间的相关性。在本问题中是指各个agent执行策略的相关性。在合作问题中换言之就是智能体之间的协调性。

## 多智能体协同方式及冲突消解方法研究

分布式系统在模拟环境下通过多智能体强化学习算法训练，智能体们可以学习到相互合作的行为策略。但实际场景中智能体的观测内容是有限的，无法掌握其他智能体的实时信息，而且实际场景中可能出现未训练过的意外场景。所以系统协作时会有行为冲突或策略重叠冗余的情况发生。基于这种场景我们设计了决策冲突消解方法。冲突消解方法主要包括冲突检测，冲突分类，以及消解过程三个步骤，而消解过程根据需求可以分为本地消解和集中消解两个方法。

* 1. 冲突检测

在战场任务被分配的时候，系统就需要为执行任务的智能体们分配冲突检测模块，一旦检测出冲突立即按照设置的冲突消解方法进行处理。检测方式需要前一节提到的决策互信模块的支持，每个智能体在战场中会实时对其他智能体的行为策略进行预测。如果推测其他智能体的行为 和自己的行为 具有重复部分，则需要判断是否属于冲突决策，并寻求解决办法。

* 1. 冲突分类

发生策略重叠情况，要立马对冲突情况进行分类。针对不同类型的应该用不同的冲突消解方法解决。

一类是效率影响型的策略冲突。例如侦察任务中，侦察机侦察路线重叠；干扰任务中，干扰机对敌方的干扰范围大面积重叠等。通过冲突消解，可以让智能体合理分配任务子目标，协助智能体合作协同，提高任务完成效率。

其次是资源节约型的策略冲突。如打击任务中，在低命中率情况下，我方智能体进行合作打击，可以有效提高命中率。但在高命中率情况下，同时发射的多枚导弹对于提升命中率的作用有限，而更有可能造成弹药资源的浪费。为了避免这种情况，需要对攻击机进行策略修正，保证我方战略资源的合理利用。

冲突分类部分的另一作用是区分智能体之间的策略重叠属于协同策略还是策略冲突。例如上述侦察任务中，若侦察机是遭受攻击临时改变路线造成的重叠，以及打击任务中，低命中率下，智能体打算通过共同攻击提高命中率等。这些受战场情况影响发生的策略变化。本地消解时，系统应认同智能体策略，集中消解时，则将情况上报给有人机。

* 1. 本地消解

本地消解是指智能体在了解到冲突情况后自行进行策略修正，解决冲突问题。

针对效率影响型的冲突，采用的是优先级配置方法解决。首先在任务分配的同时，系统会对执行任务的同机型的智能体分配优先级序列（1~n），然后在出现效率影响型的冲突时，会有限保证高优先级的智能体的策略执行，低优先级的智能体应通过预测到高优先级智能体的行动策略修改自己的任务子目标，确保任务高效完成。

针对资源节约型的策略冲突，需要根据友方命中率灵活决定。首先确定一套命中率计算系统。命中率主要由距离的决定，同时受攻击角度、高度差、设备性能等因素影响。根据仿真环境的大量训练经验，可以训练一套命中率计算方法，训练的同时可以得到一个冲突阈值，命中率高于阈值时同时攻击对于命中率提升效果差，该阈值也是冲突分类部分的计算标准。然后出现资源节约型的策略冲突时，智能体会知道包括预测友方和自己一共 n 个智能体会发起攻击，通过计算这n 个智能体的命中率，将打击任务分配给高于阈值的智能体中命中率最高的个体，自身则进行其他攻击子任务或者继续打击。

* 1. 集中消解

集中消解指由有人机处理冲突情况，需要有人机熟悉系统决策类型并及时给出统筹决策。具体过程为：

1. 策略冲突智能体需要将冲突信息同时发送给冲突对象和有人机。
2. 收到信息冲突对象和自机将处于等待调解的状态
3. 有人机收到冲突信息需要根据战场信息对冲突智能体的决策进行统筹修正
4. 等待调节的智能体在收到修正决策后根据修改后的决策行动

## 多智能体间互信机制及置信度计算方法研究

### 置信度算法实现方案

由于场景中智能体 的观测信息包括自身的观测信息 和侦察系统周期性发来的 秒之前的全局信息 。其中全局信息 中包含了其他队友智能体在 秒之前的观察信息 。而对于每个智能体，策略函数首先会根据自己当前的观察进行自身的动作决策。

其中 表示智能体 的初始决策函数。而对于其他智能体的动作预测则需要根据系统传来的全局信息 。以智能体 对智能体 的动作预测过程为例：

1. 智能体 当前的动作决策是根据 自身现在的观察得到的：；
2. 智能体 从 时刻前的全局信息分离出智能体 在 时刻前的观察信息 ，由于战场环境具有连续性，可以根据 推断智能体 当前会执行怎样的动作：

其中 表示智能体 对智能体 的策略推测函数；

1. 由于策略推测部分的输入信息具有滞后性。经过训练后 也难以做到对 的完全预测。在误差允许范围内 ，我们可以认为 对 进行了正确的预测，并且预测结果会对后续决策冲突消解环节提供根据

每次全局信息发来的时候，智能体两两之间都会互相进行决策预测。并将预测结果用于冲突消解。在此过程中预测总次数为 ，预测正确次数为 ，需要保证模型的预测准确率高于 85%。即：

基于如上设计，智能体能够给出任一智能体对其他智能体的置信度评估结果，并且能够计算任一智能体对其他智能体的置信度参数。

### 置信度方案对项目的提升意义

冲突消解模块是系统的重要组成部分，该模块的建立是基于对冲突发生之前的行为预测，从而及时避免冲突发生。

冲突场景预测功能的实现，需要每个智能体先了解本方队友接下来的行为动作，再据此判断和自己当前的决策是否存在冲突，最后根据判断结果调整决策内容。由于真实KZ场景中对抗激烈，且存在多种需紧急处置的场景，不同节点间信息传输沟通将浪费大量决策时间，这使得对友方智能体的行为预测成为系统了解友方实时决策的便捷手段。

本研究方案中的置信度概念是对智能体预测本方队友决策行为的准确程度进行量化评估，预测准确率与置信度成正相关。**通过提高智能体之间的决策置信度，可以提升冲突预测环节的准确度，从而降低智能体间出现决策冲突的概率，进而节省冲突消解时间，提升多智能体之间的协调性，最终达到提高模型效果，高效利用系统资源的目的。**

## 有无人协同互信置信度评估算法模型的开发与测试

## 面向分布式决策的模型集成及全流程仿真验证

无人机机群的仿真系统验证包括无人机数字模型设计模块，典型场景搭建设计模块，集群协同控制决策算法设计模块。

### 无人机数字模型设计

无人机数字模型设计使用的是 Xsim 仿真系统。无人机模型的设计参照实际无人战斗机的性能数值，分别完成了对无人机机体子系统、三维环境子系统、集群环境感知模块的设计实现。

### 典型场景搭建

典型场景搭建以异构空战为条件，为无人机模型提供一个虚拟空战场景，测试集群协同控制决策算法在编队行动、信息侦察、电磁干扰、协同制导、目标打击等方面的决策能力。空中作战力量运动速度快、活动空间大，考虑搭建大规模的作战场景。同时，结合单局比赛紧凑性和 AI 计算复杂度，将作战空域范围设置为 300km\*300km。战场中心为坐标原点，向右为 x 轴正轴，向上为 y 轴正轴。

作战想定为：红方作战单位包括有人机，舰载多用途无人机，无人预警机，灵巧对抗单元总计 5 类 20 个节点。蓝色方作战单位为高价值单位及其护航无人机。初始所有飞机部署均在空中，不考虑飞机的起飞、降落，所有飞机都可被 AI 控制。进行对抗空战，进攻方向为红蓝迎头进入，初始高度 9000~10000 米，速度 0.9 马赫。

初始情况下红蓝双方均了解对方兵力配置，并处于战场东西两端，需要通过侦察了解对方位置。蓝色方目标是突破红方拦截，保护高价值单位进入东部领空；红色方目标是拦截蓝色机群，打击敌方高价值目标，破坏蓝方战略目的。在此过程中，双方会进行存在干扰机协助下攻击机群的空中打击作战。

* 结算判定规则

评判仿真结束必须达到终止条件：“到达预定运行时间”或者“红方有人机全部毁伤”或者“蓝方高价值目标全部毁伤”。

1. 到达预定运行时间

达到仿真系统预定运行时间则停止仿真

1. 红方有人机全部损毁

对于红方而言编队的高价值目标就是有人机，必须全力保证有人机存活，有人单位被歼灭则仿真结束。

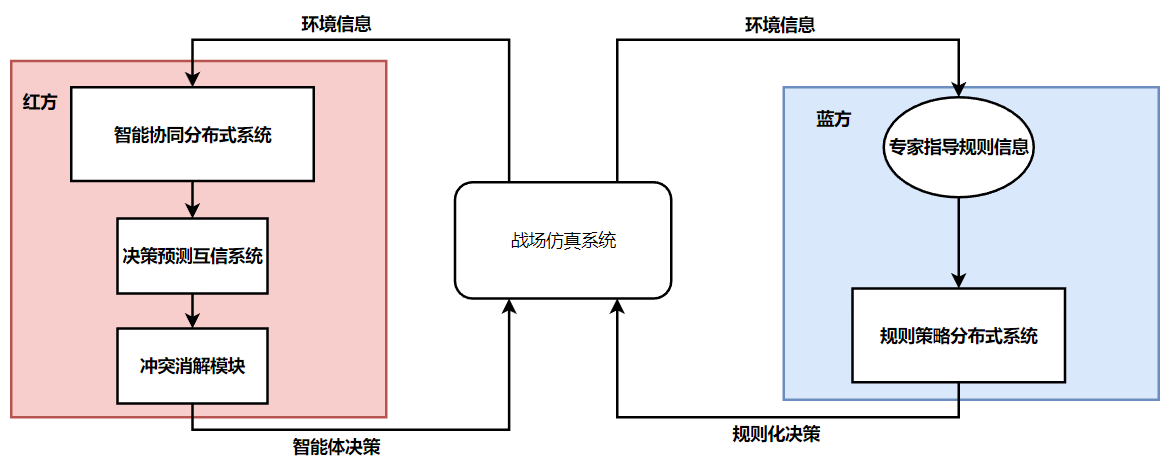
1. 蓝方高价值目标全部损毁

红方集群协同控制决策算法设计目标就是打击敌方高价值目标，完成歼敌的任务目标则仿真结束。

仿真结束系统利用仿真过程中产生的决策信息计算系统指标参数，输出仿真结果。

### 集群协同控制决策算法设计模块

本技术主要侧重于集群协同控制决策算法设计模块的仿真实现。

系统的组成结构图如下所示：

**图2.3 系统的组成结构图**

为了对分布式系统决策过程进行仿真验证，我们根据领域专家经验编写基于规则的分布式决策系统。在智能体与专家系统的空战对抗仿真中，红方采用智能分布式系决策算法进行控制，蓝方由专家系统进行控制，双方在上述搭建的仿真环境中进行空战博弈。仿真过程中收集智能分布式决策系统中每架无人机的实时的自身决策和对其他无人机的决策预测。

### 仿真结果输出

仿真结束后，首先统计 5 类 20 个节点在仿真过程中，在收到环境信息后分布式模型的响应时间。检查是否所有决策的返回时间不超过1 min时限，输出收集的响应时间的详细数据和时限检查结果。

然后分别计算单个节点 在整个仿真过程长度为的 的时间段内，每个决策时间点 上自身的决策 和其他19个节点对它的决策预测 的差距，以此统计出预测成功和预测出错的次数。最终输出置信度为预测成功次数占预测总次数的比例 。