|  |  |  |
| --- | --- | --- |
| **Problem Chosen** D | **2020 MCM/ICM Summary Sheet** | **Team Control Number** 2008495 |

Use this template to begin typing the first page (summary page) of your electronic report. This template uses a 12-point Times New Roman font. Submit your paper as an Adobe PDF electronic file (e.g. 1111111.pdf), typed in English, with a readable font of at least 12-point type.

Do not include the name of your school, advisor, or team members on this or any page.

Papers must be within the page limit specified in the problem statement.

Be sure to change the control number and problem choice above.  
You may delete these instructions as you begin to type your report here.   
  
**Follow us @COMAPMath on Twitter or COMAPCHINAOFFICIAL on Weibo for the most up to date contest information.**

**Summary**

**Have a wonderful soccer game**

本文利用图论，概率论和微积分的方法，利用数据分析和建立机器学习模型，为足球教练的阵容安排和球员训练提供策略。

Firstly，根据图论，在球员之间建立传球网络，并建立单次传球的价值评价模型，用于评价两两球员间传球的配合程度，即传球网络的边权。建立在一定时间范围内所有参与比赛的个球员的邻接矩阵，通过以个点的子完全图边权之和为排序关键字找出若干组优秀的元组合。同时建立基于时间尺度的价值模型，用于评价时间对传球效率的影响。

Secondly，我们将反映成功团队合作的绩效指标划分为静态指标和动态指标。静态指标包括球员位置安排和球队阵型（line-up），我们建立球员赛季热点模型和球员分布模型。动态指标包括opponents，side，coach，passes，defence，attack and fail等。对经过数据清洗动态指标之间通过可视化进行相关性分析后，以净胜球分类作为比赛样本标签，以随机森林分类器作为机器学习的模型，用网格搜索调优参数，建立动态指标评价模型，进行交叉验证，达到了80%的准确率。

Thirdly，球队的结构策略包括了静态指标和动态指标。通过上述中建立的模型进行观察分析，我们着重研究静态指标对球队的胜利起到的关键作用，综合考虑球员位置和技术数据评价模型，建立不同球员在不同位置价值评价模型。通过模拟退火算法，优化11人排列组合的考虑，在局部最优解的父级搜索树进行搜索全局最优解，最终给出价值最优的首发阵容阵型图。此外我们还考虑以下三个次要影响因素：球员间默契度，主客场影响和教练安排并给教练提出综合建议。

Finally，我们用哈士奇球队的案例来解释群体动力学。并用哈士奇球队建立模型得到的结论来说明如何设计更有效的团队，并对团队绩效指标进行补充。

**Contents**

**1.Introduction 1**

1.1 Background 2

1.2 Our work 3

**2.Problem restatement 4**

**3.Terminology 1**

3.1 Terms 2

3.2 Symbols 3

**4.Preparation of the Models 4**

4.1 Assumptions 5

4.2 Notations 6

**5.The Models 1**

5.1 Model 1 2

第3级 3

5.2 Model 2 2

第3级 3

5.3 Model 3 2

第3级 3

5.4 Model 4 2

第3级 3

**6.Strengths and Weaknesses 4**

6.1 Strengths 5

6.2 Weaknesses 6

**References**

# 1.Introduction

## 1.1 Background

足球运动有着悠久的历史，从推广以来就受到了全世界的喜爱，球迷遍布全世界。足球可以说是当今世界普及度最高的运动项目之一，足球这看似简单的一种运动，背后蕴藏着个人能力及团队协作的奥秘。随着时代的发展和科技的进步，足球运动员们和教练在技术上不断提升，给观众们呈现出一场场精彩纷呈的比赛。众所周知，一场精彩的足球赛离不开球员和球队贡献的，通过研究团队中每个人的行动，平衡队内力量，协调团队关系，合理安排上场时间和每个人所在的位置，可以在最大程度上得分。

## 1.2 Our work

我们的工作是从传统模型开始，在问题一中，我们首先建立每一场中十四位球员（其中包括三位后补球员）的传球次数统计图，再建立整个赛季中

我们的任务时为足球教练提供战略决策，在千变万化的球场上找出一些较为普遍适用在作战规律

在建立模型时，我们使用了Python，vscode，OneNote（开始胡编乱造，要删掉）

# 2.Problem restatement

**2.Problem restatement**

足球赛是一项老少皆宜的运动，自从其纳入国际赛事以来，人们就创造出各种各样的方法来评价整个比赛和整个赛季的团队动态，来帮助确定下个赛季可以改善团队合作的具体策略。我们需要使用ICM团队提供的数据建立模型来解决以下四个问题。

1.将每一个球员当做一个节点，创建传球网络来识别二元配置，三元配置和多元配置。我们需要建立在传球网络下，单次传球的价值评价模型，以及时间结构指标的传球总数评价模型。

2.确定反映成功团队合作的绩效指标，我们需要考虑静态指标和动态指标。建立每个绩效指标对成功团队合作影响的模型，并用一个模型来囊括这两个子模型。

3.通过对问题1,2中建立的模型的观察分析，告诉教练什么样的结构策略适用于哈士奇球队。用模型分析的结果为教练提高球队的下个赛季的成功率给出建议。

4.用哈士奇球队的案例来解释群体动力学理论，用哈士奇球队建立模型得到的结论来说明如何设计更有效的团队，并对团队绩效指标进行补充。

# 4.Preparation of the Models

# 5.The Models

## 5.1 model 1

为了构造结构化的传球网络，用来分析球员间传球配合的默契程度，应在不同多维度、状态变化情况下进行分析。例如从微观上两两球员之间的行为，到宏观上多个球员之间的行为；以及时间尺度从比赛中的单位时刻到整个赛季。

### 5.1.1 传球评价Index（PEI）

两人传球评价指数，用于评价两人配合程度。在一场球赛中，宏观来看，球员相对于球场可以视为一个节点，球场可以视为一个网络，每一次传球可以视为节点之间的连线。以两人之间每次传球累计评价作为2人的传球评价指数。多人传球网络体系中在传球评价体系中，将三个节点连接成闭合环路，边权之和即为3人传球评价指数。

通过生活经验和数据挖掘所发现的规律可以构建PEI计算模型：

1. 传球类型权重表：

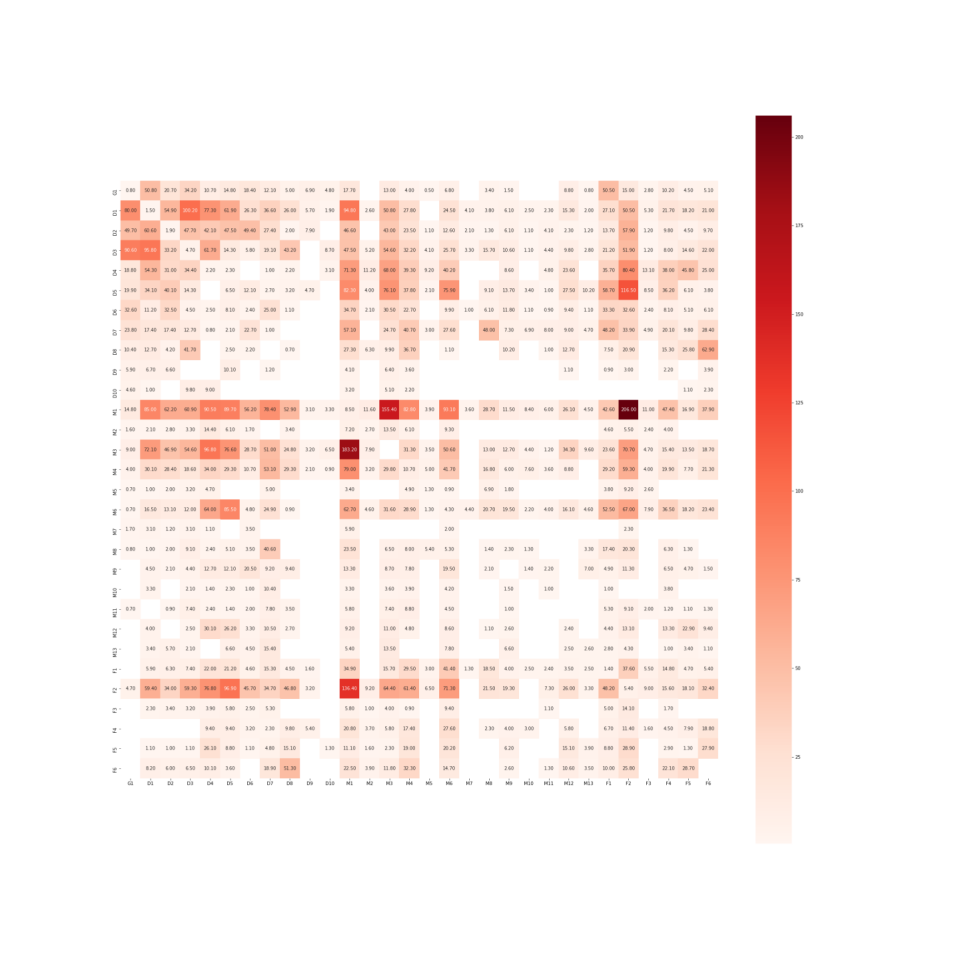


1. 计算传球或接球时分别受到的防守压力

其中，x为球员到对方球门的横坐标，与受到的防守压力成负相关

1. 两人之间的边权，即单次传球评价，为此次传球类型的权重乘以防守压力加权平均数，量化为以下公式。其中，为传球人坐标带入得到的值，为接球人坐标带入得到的值。

根据这一传球评价指数模型，统计出每场球赛中每两位球员间传球的次数并绘制出38组14×14的矩阵数据和代表整个赛季的30×30的矩阵数据。由得每两人之间所有传球价值评价总和图：



整个赛季单次评价图

### 5.1.2 传球网络模型构建及识别网络模式

网络上两两球员之间的联系，宏观上体现为球员间传球的评价总和。筛选两人传球评价超过一定阈值的边，运用图论的方法选择性剔除交叉边，将基于传球评价指数构建的line-up传球网络可视化，用线的深浅表示：



单场比赛的两两传球网络图

从这一模型的可视化中，我们可以直观地分析传球配合频繁和默契地球员，还可以直观的看出主力球员中多人传球配合的组合。（列出3人组、4人组的结果）

### 5.1.3 时间尺度上传球状态波动

传球状态会在时间尺度上进行波动。定义

并以作为球队实时状态的指标。球赛刚开始时，球员身体还未warm up，导致传球概率密度较小，5—10分钟过后，传球效率逐渐提高并大致趋于稳定，即：

As the time goes，球员们体力消耗，传球密度降低，即传球数量的增速减缓（虽然一场球赛中成功传球次数仍在增加，但传球失败概率开始增加），此后传球密度呈现下降趋势，即：

纵观整个赛季中38场球赛球员们的传球概率密度，整体上看可以与单场球赛所展示出的概率密度变化趋势相同。令时间为横坐标，成功传球密度为纵坐标作图：



第1场比赛传球密度 整个赛季传球密度

总体来说，传球的密度在时间尺度上相对稳定。若使用Monte Carlo method对每次传球进行模拟，设定在上一次传球后下一次传球还需要的时间概率分布服从，其中为统计的平均传球间隔时间，则在样本规模满足时会近似与左图的分布；随着样本规模增加，当满足后则会近似与右图的分布。因此我们可以认为每一个时间点发生传球事件的概率服从。

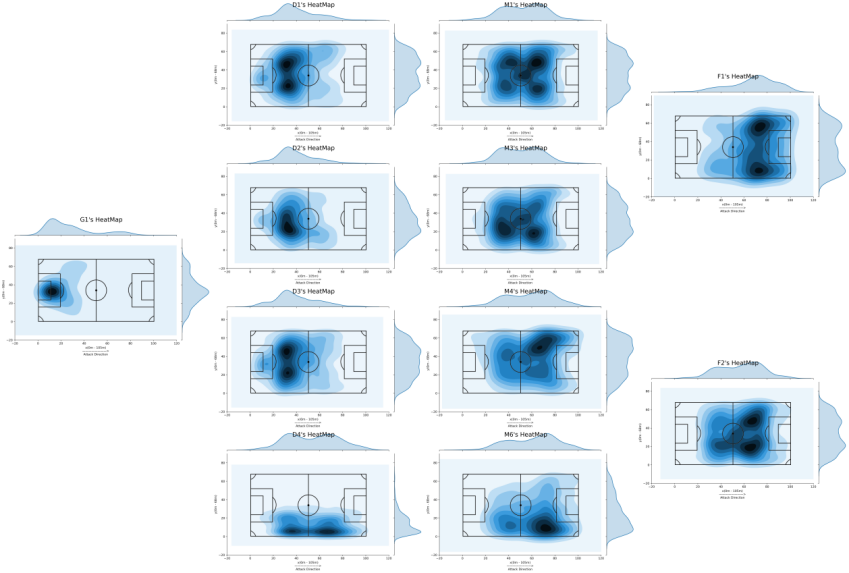
## 5.2 model 2

足球队中成功团队合作有许多指标，通过数据分析和实际经验，我们主要考虑以下indicators：静态指标和动态指标。首先，我们使用 评价一场比赛的球队整体发挥，作为单场比赛表现标签，定义：

### 5.2.1静态指标

为了考虑球员位置分布，我们采出每个球员在整个赛季中的位置坐标，做出球员运动位置的热点图，热力图每个点的值定义如下：

颜色越深则表示出现在此处的概率较大，越浅表示出现的概率越小。经过的计算，主力11人的位置热点图如下：



球员位置分布热点图

在一场球赛中，球队的阵型对团队协作起到重要作用，我们考虑在一场球赛中球员阵型，我们采取每一场比赛中每一位球员的运动坐标，采用坐标对时间积分的方法，找出每场球赛中，每一位球员平均坐标。将在数据中可以获取（球员出现在Origin/Destination）的时间点作为新的横坐标，X或Y坐标作为新的纵坐标，得出函数。我们近似认为在任意两个有记录的时间点，球员在X或Y方向上匀速移动，这样就将离散型的数据集转换为了连续性的数据集。因此平均坐标，以X坐标为例，Y坐标同理：

将这11位球员的位置标在图中绘制出每场球赛的阵型图，部分阵型图如下：



Match 1 and Match 11球员阵型图

### 5.2.2动态指标

动态指标包括了球队人为影响因素和在比赛里产生的技术数据：人为影响因素包括了教练、对手水平、主客场，技术数据包括了射门、传球、解围在内的各种events统计。原始的数据以单个事件作为样本的单位，而我们将其分类统计为以一场比赛为单位的动态类型数据，通过观察以新结构存储的数据，提取出其中的若干特征信息。

#### 5.2.2.1数据清洗和特征工程

In feature engineering，为了降低特征的维度，不仅使用PCA筛选并剔除影响不显著的特征，还可以使用ChiMerge这一特征分箱的方法，将EventSubTypes分为传球，进攻，防守和Fail四个方面，与教练、主客场、对手水平一起作为一场比赛的特征。通过标准化、哑变量、结合分析等方法处理统计后的数据来量化比赛的特征：

（1）统计型数据

（2）多事件结合分析型数据

（3）One-Hot编码哑变量数据

#### 5.2.2.2 可视化分析

分析对于影响：



主客场与得分的关系图

时的分布更多，分布更高，因此主场表现结果整体上比客场要好。

分析不同Coach的执教水平以及对于球队的指导成效：





不同教练指导下球队4种表现数据和净胜球对比图

从boxen图我们可以看出，在Coach 3指导下，球队等数据较好，其次是Coach 2和Coach 1。我们还可以得出教练们的执教风格，例如：教练1更具侵略性，防守就显得平庸；教练2强调强硬防守；教练3则较为平衡，战绩最佳。

分析、对于的贡献：



进攻、传球与净胜球之间关系图

从图中我们可以看出，在不同净胜球数下，进攻和传球大体上为线性相关，斜率为正。

与呈正相关，且分布越集中，的方差较小。我们可以得出结论：在一场球赛乃至整个赛季，越多，大概率有着更高的。

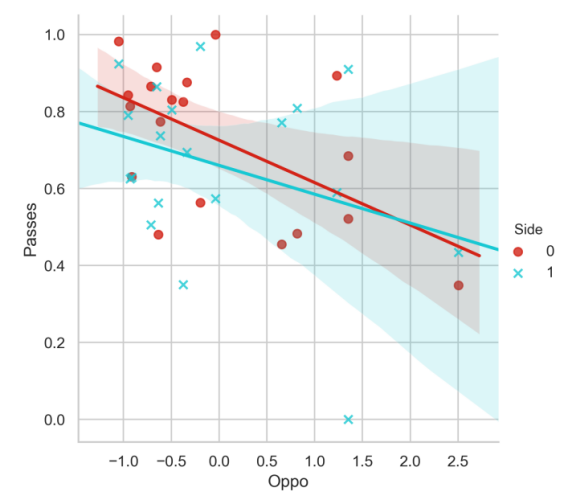
分析、对于的贡献：



防守、失败与净胜球之间关系图

与呈正相关，与呈负相关，且分布越集中，的方差较小。观察发现：图2左1的点分布在下方，因此防守不好会导致输球；右1左半边没有点，因此期望赢球则失误不能多。

以作为考察球队整体表现的positive指标，结合指标进行多角度分析：



左：进攻和防守之间关系图 右：对手水平与传球关系图

从左图中我们可以看出数据重心分布在右下角，认为整个赛季上（进攻表现）显著优于（防守表现）。从右图中我们可以看出不论是在主场还是客场，，但主场更可能有较小提升；结论是对手水平越高，我方传球率越低。

综合所有处理得到的特征，通过Pearson相关系数的计算来估计出变量间两两特征相关性。

令矩阵，得：



动态指标整合的特征相关系数矩阵

#### 5.2.2.3 model建立and训练

我们以作为每场比赛评价标签，希望学习后的模型能够基于处理后的数据对比赛进行分类，对应到的标签。由于个特征数量较多，且与标签相关性不一，不宜采用线性模型进行分类；且样本数据数量极少，在尝试一些深度学习算法时容易有较大偏差。综上，我们选择随机森林模型建立标签分类器。

随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定。对于很多种资料，它可以产生高准确度的分类器；它可以在决定类别时，评估变数的重要性；在建造森林时，它可以在内部对于一般化后的误差产生不偏差的估计。建立随机森林分类器Random Forest Classifier的方法如下：

1. 输入特征数目，用于确定决策树上一个节点的决策结果；
2. 利用Bootstrap取样，从个训练用例中以有放回抽样的方式，取样次，形成一个训练集，并用未抽到的用例作预测，评估其误差；
3. 对于每一个节点，随机选择m个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这m个特征，计算其最佳的分裂方式；
4. 每棵树都会完整成长而不会剪枝，这有可能在建完一棵正常树状分类器后会被采用。

随机森林分类器的训练后，使用网格搜索grid search进行参数调优，选定

作为参数，利用K折交叉验证验计算其accuracy score，用于评估模型准确率。

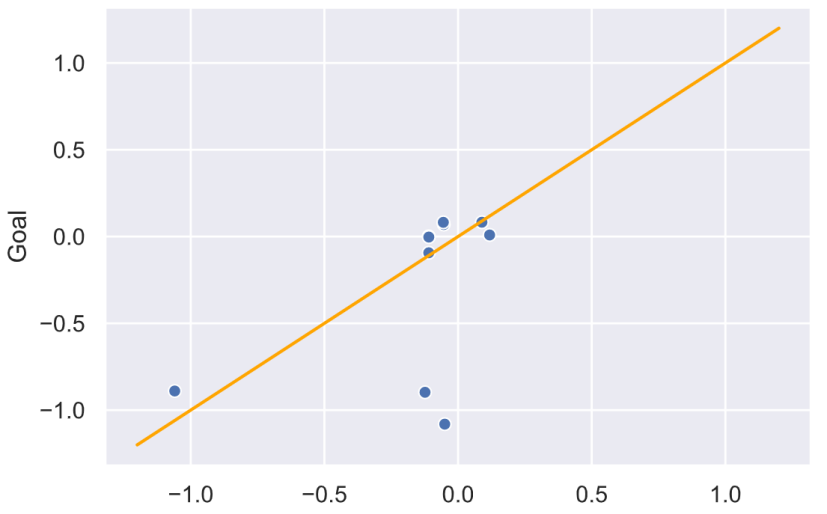


图 1模型预测交叉验证情况

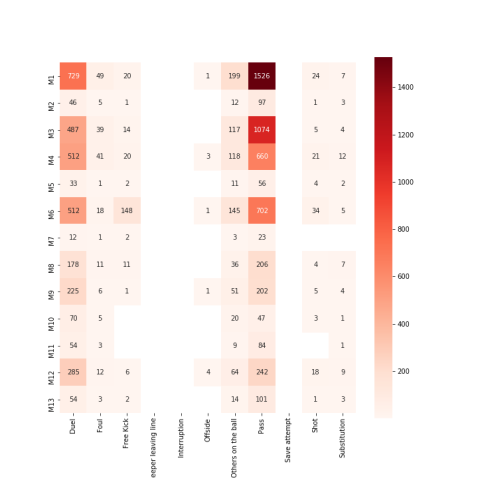
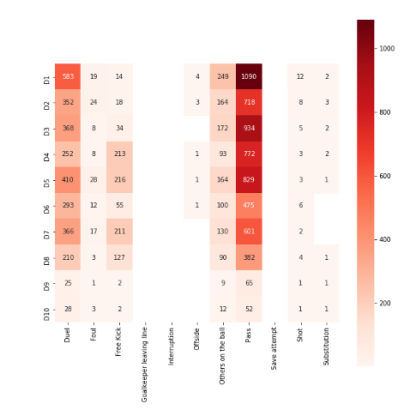
经过一定的数据调整和多次模拟结果，平均情况下得分为，最好的数据情况下可以达到的得分，在样本规模仅有的情况下，我们可以接受这一模型通过动态指标对比赛净胜球情况进行预测的准确率。

## 5.3 model 3

结构策略影响着成功的团队合作，作为一名成功的教练应该具备较好的统筹规划，协调合作，人员安排能力。我们认为，具体的结构策略应该最主要体现在以下两个个方面：球员位置安排和球队阵型。此外，还应考虑球员间默契度，主客场影响，教练安排。

### 5.3.1 位置evaluation engineering

在考虑球员位置安排时，需要计算守门员，前锋，中场，后卫四个位置不同球员的贡献值。我们在数据集中采集哈士奇球队30位球员的EventType，并以其为横轴，球员编号为纵轴，统计出每个球员在整个赛季中每个EventType次数，用颜色的深浅来表示次数的多少，以下分别为前锋，中场和后卫的EventTypes统计图：

前锋&中场&后卫EventTypes统计图

由以上四个图，我们可以看出F中贡献最大的是F2，其次是F1,F6,F5,F4。M中贡献最大的是M1，其次是M3,M4,M6。D中贡献最大的是D1，其次是D3,D5,D4,D2,D7,D6,D8。

我们希望能有实际的模型来对不同球员球员在不同位置表现进行量化评价。此时需要结合实际知识，分析不同位置各自的重要数据，通过不同EventTypes权重分配、结合球员各种能力performance，进行计算，作为evaluation of球队的29位球员（除守门员）分别在G,F,M位置上的表现情况。下图中，颜色越红表示越适合这个位置，反之越蓝则表示越不适合。



不同球员在不同位置评分表

### 5.3.2 基于SA算法优化排列组合

我们分析整个赛季38场比赛中主力阵容/首发阵容line-up，希望建立模型为教练建议最好的球队阵容该模型的目标是要找到一个最优的有序组合，使场上11人在各自位置的能力之和最大。把场上11个位置有序排列，用11位的30进制格雷码表示当前状态；例如格雷码0A1GRD739KI表示11个位置依次有第0,10,1,16,26,13,7,3,9,11,18号球员。在搜索树极为庞大、算力资源有限的情况下，我们选择模拟退火算法。模拟退火算法主要优点之一就是能以一定的概率接受目标函数值不太好的状态，且在迭代的过程中不断能够接受使目标函数向好的方向前进的解。模拟退火算法的具体步骤如下：

1. 给定冷却进度表参数及迭代初始解.以及,其中冷却进度表参数包括：控制参数的初值，衰减函数，终值以及链长度；
2. 参数时，按照如下过程作次试探搜索：
   1. 根据当前解的性质，产生一个随机偏移量，从而得到一个当前解邻域的新的试探点；
   2. 产生一个在区间上均匀分布的随机数,计算出在给定当前迭代点和温度下与接受准则相对应的转移概率：
   3. 试探搜索小于次，返回步骤1，否则进入步骤3；
3. 根据给定的温度衰减函数产生新的温度控制参数，及链长度，转入步骤2，进入下一温度点的平衡点寻优。



图 2 SA Flowchart

在实际试探搜索中，我们很可能现入局部最优，需要进行判定以退出。当前解的优化程度小于当前最优解的优化程度的时候,新解被接受的概率为,而当温度足够低的时候，较差解被接受的概率趋近于。依据最近的次搜索中都没有优化程度更高的解出现这一特征，可以根据具体问题确定阐值而后判定搜索己经进入局部最优。

### 5.3.3 其他结构策略因素

考虑完主要策略后，我们考虑以下三个次要影响因素：球员间默契度，主客场影响和教练安排。

首先，选择默契度高的小分队有利于提高传球和进球的效率。默契度高的小组往往配合能力较强，有助于比赛的成功。传球效率较高的球员往往适应能力较强，与其他球员的配合度也较好。

主客场因素也是必须要考虑的，有的球员适应性较强，在主场和客场都能较好的发挥出原有的水平，而有些适应性较差的球员只在主场发挥出原有的水平，环境对他的表现有较大的影响。那么在主客场时，应该安排不同的球员上场。

最后，教练安排上，整个赛季中，Coach 1,Coach 2,Coach 3,分别指导了9,5,24场比赛，通过我们在第二题中的数据分析也可得出，Coach 3的水平较高。

### 5.3.4 structural strategy conclusion

纵观整个模型，为了在下个赛季中提高球队成功率，我们团队给出的建议是，球队聘用Coach 3作为球队主教练，采用442的line-up，将F1,F6,F2,M3,M1,M6,D3,D1,D2,D5作为主力球员，命名该阵容为，他们的位置按下图安排：



球员阵型图

其中的一个特殊点，根据赛季数据的评价，F2球员作为一名前锋，拥有很强中场的能力，在尝试将他安排在中场时取得了显著的新最优解，这说明每个人的任一位置评价较为重要，侧面indicate我们的模型因素考虑完善。

综上，该阵型的个人能力总分为，团队配合得分为，按照进行加权平均，最后得到综合评分为。实际比赛中与此相似的阵型取得很好的战果，也验了我们的评价模型和模拟退火算法的可行性和准确性。

## 5.4 model 4

团体动力学主要包括五个方面的内容:团体内聚力，群体压力和群体标准，个人动机和群体目标，领导与群体性能以及群体的结构性。

群体内聚力可以指球队想赢的信念，群体压力是来自外界的压力。信念是来自球员内心的自我要求，压力是来自外界的威胁。提高群体内聚力是提高球队赢球信念的良性循环，见效慢但是正反馈。而群体压力的是假要适当，否则会将人压垮。

群体内聚力偏向于给予球员精神支柱。责任性行为体现在每个球员各司其职，踢好自己的位置。成员之间的相互影响表现在球员间的相互鼓励，相互进步。价值取向的一致性是指球员对球队的发展战略一致认同，不仅是对临场战术的认同，还可能是对球员更替的认同。成员安全感的发展即球员对球队的信任，对球队有归属感。群体生产力的提高就是球队的进步，可以包括技术上和战术上的进步。

群体标准有很多，宏观来看，有对球队战绩的指标，微观来看，有球员间的相互要求。群体标准可以帮助球队和球员更有压迫感，相互竞争是提高个人能力，更加集中于打好每一场比赛，由此观之，适当的群体压力是有必要的，有竞争才会有发展。

对于球队来说，群体目标就是赢球，球员的行为是各司其职。群体目标影响群体行为，当球队的目标和球员的目标达成一致时，球员会表现出最为强烈的求胜动机，努力为进球而拼搏。

在球队中，领导者有两种，可以是为球员提供训练，赛前指导和赛后总结的教练，也可以是在赛场上鼓舞士气，树立榜样和调度指挥的队长。领导者的能力会影响球队的发挥和球队的进步。此外，教练和队长对球员的关照可以很好地鼓提高队内活力。

当一个球队在球员关系的安排上获得稳定时，拥有了一定的球队结构。稳定的结构有利于培养球员间默契，从而有更大的概率赢得球赛。