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

Summary

Have a wonderful soccer match

This paper proposes a method, with graph theory, probability theory and calculus, to build machine learning models based on data analysis, which aims at providing strategies for soccer coach's lineup arrangement and players' training.

Firstly, the Pass Network Model can be established according to the graph theory, whose edge-weights are evaluation of coordination degree of each dyadic configurations. Pass Evaluate Index is designed for evaluate a single pass, and the summation of each pass can be defined as the edge-weights of PNM. For analysis, the adjacency matrix of N participating players within a period. Several outstanding M configurations can be found by the sort of M-element combination with the key of the sum of the sub-complete graph edge weights. What's more, investigation of the influence of time on Pass Density depends on the constructed and approximate function of time and pass.

Secondly, performance indicators that reflect successful teamwork can be divided into dynamic indicators and static indicators. Static indicators include player position arrangement and line-up with which player season heatmap models and player position models can be established while the dynamic indicators include opponents' strength, side, coach, passes, defense, attack and fail. etc. After visualized analysis of the correlation between the dynamic indicators extracted after data cleaning, and with the setting label by the goal difference, the random forest classifier is used as the machine learning model as the evaluation model of dynamic indicators. After the Grid Search used for tuning parameters, and cross-validation, the accuracy of the model achieving 80% approximately.

Thirdly, the study focuses on the role of static indicators in the performance of the team and establishes different players' value evaluation models in different positions which comprehensively consider the player's position and technical statistical data evaluation. To optimize the value of 11-person permutation, we choose simulated annealing (SA) algorithm which searches the global optimal solution in the cousin points in the same minimized search tree as the local optimal solution has attained. The model finally gave the best starting lineup formation. In addition, we also consider the following three secondary factors: tacit understanding between players, home and away influence, and coaching arrangements. All analysis above can be concluded as comprehensive suggestion to the coach.

Finally, we use the case of the Huskies to explain group dynamics. And use the conclusions obtained by the Huskies to build a model to explain how to design a more effective team and supplement the team performance indicators.

Key words: Network; Graph theory; Random forest classifier; Simulated annealing; Heat map; Group dynamics

Contest

1 Introduction

1.1 Background

Football has a long history. It has been loved all over the world since it was popularized. Football can be considered as the most popular sports in the world. Football, a seemingly simple sport, contains the secrets of individual ability and team cooperation. With the development of the times and the progress of science and technology, football players and coaches continue to improve in skills, showing the audience wonderful matches. As we all know, a wonderful football match is inseparable from the contributions of players and teams. By studying the actions of everyone in the team, coordinating the team relationship, reasonably arranging the minutes and line-up, we can score best.

1.2 Problem restatement

Football is a sport suitable for all ages. Since its inclusion in international tournaments, people have created a variety of methods to evaluate the team dynamics throughout the match and over the entire season to help determine specific strategies that can improve teamwork next season. We need to use the data provided by the ICM team to build a model to solve the following four problems.

1. Consider each player as a node and create a passing network to identify dyadic, triadic and multiple configurations. We need to establish a value evaluation model of a single pass and a general evaluation model of the passing of the time structure index under the passing network.
2. To Identify performance indicators that reflect successful teamwork, we need to consider static and dynamic indicators. Establish a model of the impact of each performance indicator on successful teamwork, and use one model to encompass these four sub-models.
3. By observing and analyzing the model established in Questions 1 and 2, tell the coach that which form of structural strategy is applicable to the Huskies. Using the results of the model analysis to make suggestions for the coach to improve the team's success rate next season.
4. Use the case of the Huskies to explain the theory of group dynamics, and use the conclusion of the model established by the Huskies to explain how to design a more effective team, and supplement the team performance indicators.

2 Preparation of the Models

2.1 Processing Tools

Tool	Uses
Visual Studio Code 1.42	Coding, Visualization
IPython 3.6.8	Run Code
Visio	Design Flowchart
Excel	Arrange Dataset
GitHub	Synchronization, Storing
MindMaster	Plot Mind Map

Table 1 Tools

2.2 Data Cleaning

Data Name	Processing Type	Feature Name
Side	Map + Dummy	Side_1, Side_0
Coach	Dummy	Coach_1, Coach_2, Coach_3
Opponent Strength	Analysis	Oppo
Shots	Count	Attack
Dribbles		
Touch		
Corner		
Offside		
Tackle	Count	Defence
Dispossess		
Aerial Won		
Interception		
Clearance		
Blocks		
Saves		
Passes	Count	Pass
Possession	Search + Integrate	
Pass Success	Calculate	
Foul	Count	Fail
Loss of Possession	Search + Count	

Table 2 Data Processing Method

3 Establishment of PNM and Analysis of Influence Factors

In order to construct a structured passing network, which is used to analyze the tacit understanding of passing between players, it should be analyzed in different dimensions and states. For example, from the behavior between two players at the micro level, to the behavior between multiple players at the macro level; and the time scale from the unit time in the match to the entire season.

3.1 Pass Evaluation Index（PEI）

The evaluation index of pass between two players $PassValue(p_i, p_j)$is used to evaluate the degree of cooperation between them. In a match, from a macro perspective, players can be regarded as nodes, the field can be considered as a network, and each pass can be considered as the connection between the nodes. We define $PassValue(\ _i, p_j)$ as the pass evaluation index for each pass. In a multiplayer pass evaluation system, three nodes are connected into a closed loop, and the sum of edge weights is the 3-player pass evaluation index.

$$PassValue(p_i, p_j, p_k) = EdgeValue(i,j) + EdgeValue(i,k) + EdgeValue(j,k)$$

According to the experience of life and the rules discovered by data mining, a PEI calculation model can be constructed as follows:

(1) Weight table of pass types:

$$\alpha_i \ is \ constant \ for \ i \ in \ Pass \ Types.$$

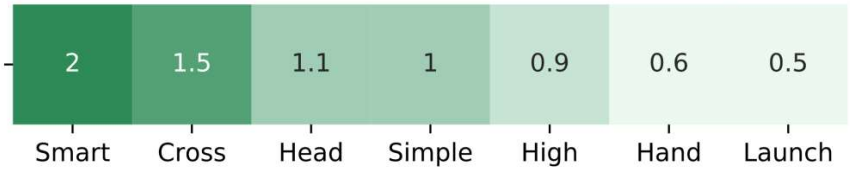

Figure 1 Pass weights

(2) Calculate pressure from defenders when passing or receiving

$$DefPress(p_i) = 1 - \frac{1}{2}\tan[\frac{11}{6} \times (\frac{x_{p_i}}{100} - 0.6)]$$

For x is the abscissa from the player to the opponent's gate, and it is negatively related to the pressure from defenders.

（3）Single pass evaluation, $EdgeValue(i,j)$, is the weight of the pass type multiplied by weighted average of the pressure from defenders, quantified as the following formula:

$$\begin{cases} EdgeValue(i,j) = PassValue(p_i, p_j) \\ PassValue(p_i, p_j) = \sum_{i\ in\ Pass} \alpha_i * \left(DefPress(\ _{i-from}) * 0.3 + DefPress(p_{i-to}) * 0.7\right) \end{cases}$$

According to this pass evaluation index model, an adjacency matrix Arr of all N players participating in the match within a certain time range is calculated. From $Arr[i,j] = PassValue(p_i, p_j)$, we can get the sum of all values of pass evaluations between each two players:

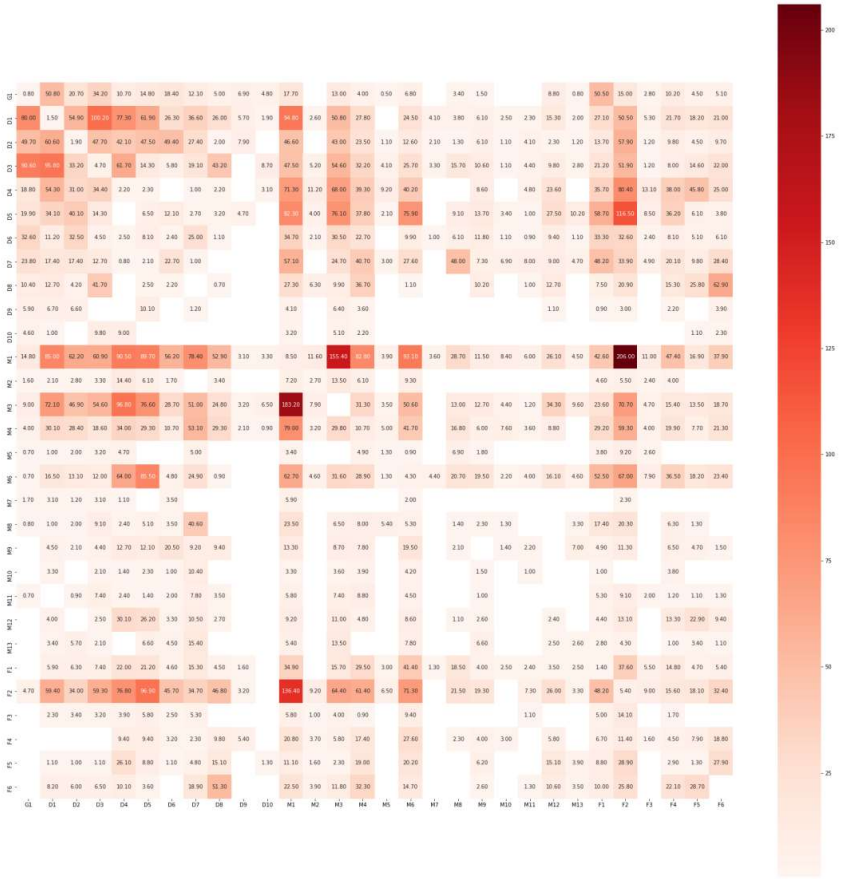

Figure 2 PassValue(i) for whole season

3.2 Pass Network Model (PNM) and Recognition of Network Pattern

The connection between two players on the network, Macroscopically, it is the sum of the evaluation of passing between players. Screening the edges whose pass evaluation exceeds a certain threshold, using graph theory methods to selectively remove the crossing edges, and visualize the line-up passing network constructed based on the pass evaluation index, which is expressed by the shades of the line.

$PassValue(\ _i, p_j)$:

$$\begin{cases} lighter, PassValue(p_i, p_j)\ is\ less \\ deeper, PassValue(\ _i, p_j) is\ more \end{cases}$$

$$Color_{Pass}(i,j) = Palette\left(\frac{PassValue(p_i,p_j) - \min\limits_{\substack{a\neq b \\ \{a,b\ in\ players}} PassValue(p_a,p_b)}{\max\limits_{\substack{a\neq b \\ \{a,b\ in\ players}} PassValue(p_a,p_b) - \min\limits_{\substack{a\neq b \\ \{a,b\ in\ players}} PassValue(p_a,p_b)}\right)$$

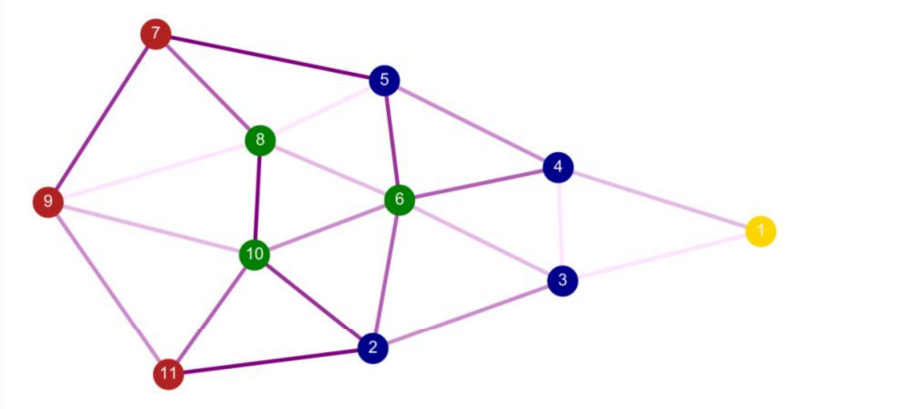

Figure 3 Pass network in one match

From the visualization of this model, we can intuitively analyze the players who pass frequently and tacitly, and can also see the combination of multiple passing cooperation among the main players intuitively.

N	Players				Score
2	M1	F2			342.4
	M1	M3			338.6
	D5	F2			213.4
3	M1	M3	F2		816.1
	D5	M1	F2		727.8
4	D5	M1	M3	F2	1113.5

Figure 3 Players configuration

3.3 Fluctuation of Passing State at the time

Passing frequency fluctuate on the time scale. $Passes(0,t)$

$$Passes(0,t)\ is\ the\ sum\ of\ passes\ before\ t.$$

And $Passes(0,t)$ is used as an indicator of the team's real-time status. At the beginning of the match, the players' bodies were not warmed up, resulting in a low probability of passing. After 5-10 minutes, Pass efficiency gradually improved and generally stabilized, that is:

$$\begin{cases} Passe\ '(0,t) > 0 \\ Pass\ \ ''(0,t) < 0 \\ \lim\limits_{t\to t_0} Passes\,(0,t) = P_0 \\ t_0 < \dfrac{Halftime}{2} \end{cases}$$

As the time goes, the players' physical strength and the pass density decrease, that is, the increase in the number of passes slows down (although the number of successful passes in a match is still increasing, the frequency of pass failure begins to increase), after that the pass density Showing a downward trend, that is:

$$\begin{cases} Passes'(0,t) < 0 \\ Passes''(0,t) > 0 \\ \lim\limits_{t\to t_0} Passes\,(0,t) = P_1 < P_0 \\ t_0 > \dfrac{Halftime}{2} \end{cases}$$

Looking at the pass frequency density of players in 38 matches throughout the season, one the whole the trend of frequency density is the same as that showes in a single match. Plot with time as the abscissa and successful pass density as the ordinate:

$$Passes(t_1,t_2) = \int_{t_2}^{t_1} PassDiv(t)dt$$

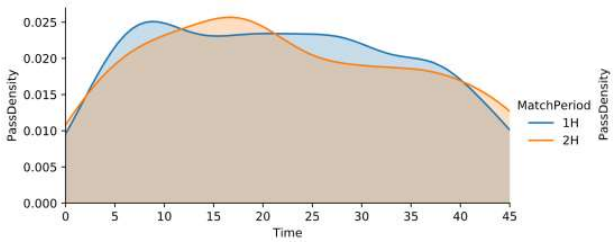

Figure 4 Whole Season Pass Density

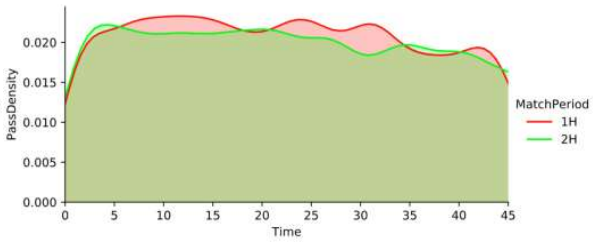

Figure 5 No.1 Match Pass Density

Generally, the density of passes is relatively stable in a time scale. If we use the Monte Carlo method to simulate each pass, the time probability distribution for the next pass after the last pass is set to obey $N(0,1)$, where θ is the statistical average of interval time of pass. Then when the sample size N satisfies $\log_{10} N \approx 2$, it will approximate the distribution on the left graph; as the sample size increases, when $\log_{10} N > 4$ is satisfied, it will approximate the distribution on the right graph. Therefore, we can consider that the probability of passing events at each time point obeys $N(0,1)$.

4 Soccer Team Indexes and Performance Prediction Based on ML

There are many indicators for successful teamwork in a football team. Based on data analysis and practical experience, we mainly consider the following indicators: static indicators and dynamic indicators. First, we use $Goal(G_i)$ to evaluate the overall performance of a team in a match. define $Goal(G_i)$:

$$Goal(G_i) = \begin{cases} -1, & OwnScore - OpponentScore < -1 \\ 0, & OwnScore - OpponentScore \in [-1,1] \\ 1, & OwnScore - OpponentScore > 1 \end{cases}$$

4.1 Static Index (SI)

In order to consider the distribution of players' positions, we took the position coordinates of each player throughout the season and made a heat map. The value of each point in the heat map is defined as follows:

$$Heatmap_{p_k}[i,j] = \frac{1}{4\delta^2}\int_{x-\delta}^{x+\delta}\int_{y-\delta}^{y+\delta}\begin{cases}1, player\ has\ been\ here\\0, player\ never\ passed\end{cases}dxdy\ ,\delta > 0$$

The darker the color is, the more active the player is in this position. After calculating $Heatmap_{p_k}[i,j]$, the position heat map of the main 11 players is got as follows:

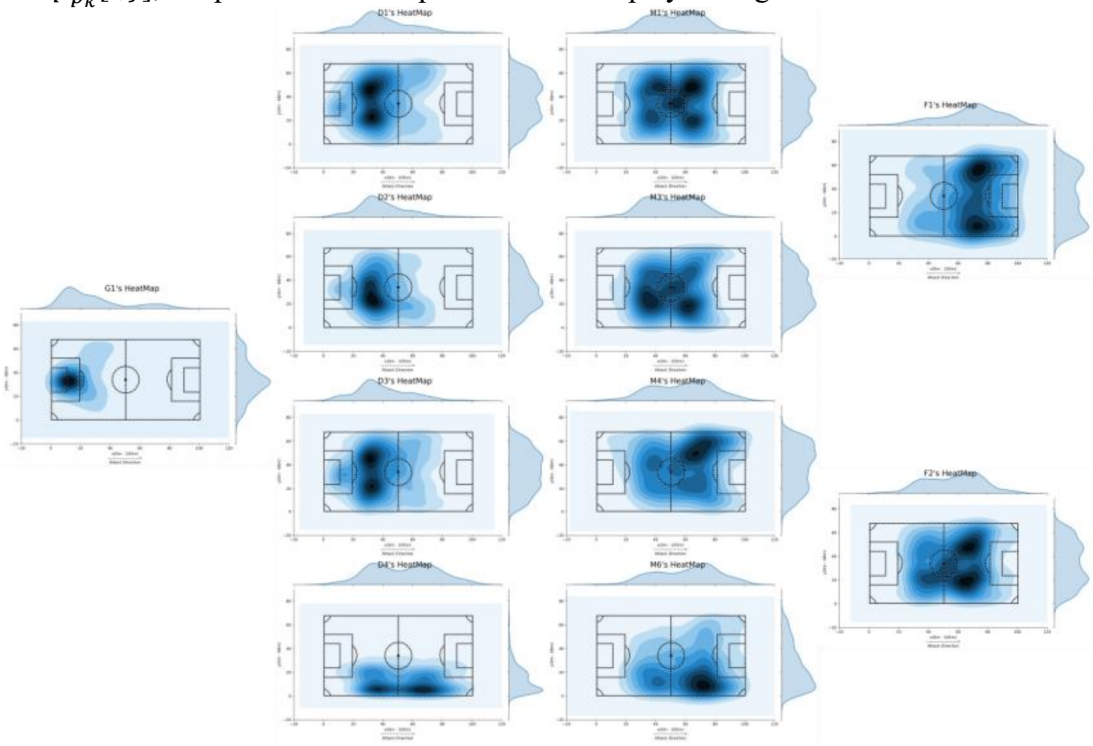

Figure 6 Players Heatmaps

In a match, team formation plays an important role in collaboration. We want to find out what the formation is like. We take the coordinates of each player in each match and integrate the coordinates over time to find out the average coordinates. We take the time which can be got from data (Origin / Destination Player) as the new abscissa, and the X or Y coordinate as the new ordinate, so we got functions X (t) and Y (t). Approximately, we thought that between any two closest recorded time points, the player moves at a constant speed in the X or Y direction, so that the discrete data set is converted into a continuous dataset for each match. So the average coordinate, taking the X coordinate as an example (the Y coordinate is the same), is:

$$X(t)\ is\ a\ piecewise\ function, X_t\ is\ the\ X\ exactly\ when\ t.$$

$$\begin{cases} AvgX(p_i) = \int_0^{90min} X(t)dt \approx \sum_{i=1}^{n}\left[\frac{1}{2}(t_{i+1} - t_{i-1}) \times X_t\right] \\ n = num\ of\ our\ events \end{cases}$$

Plot these 11 players' average coordinate on the map, we got the formation graph of each match. Some of them are as follows:

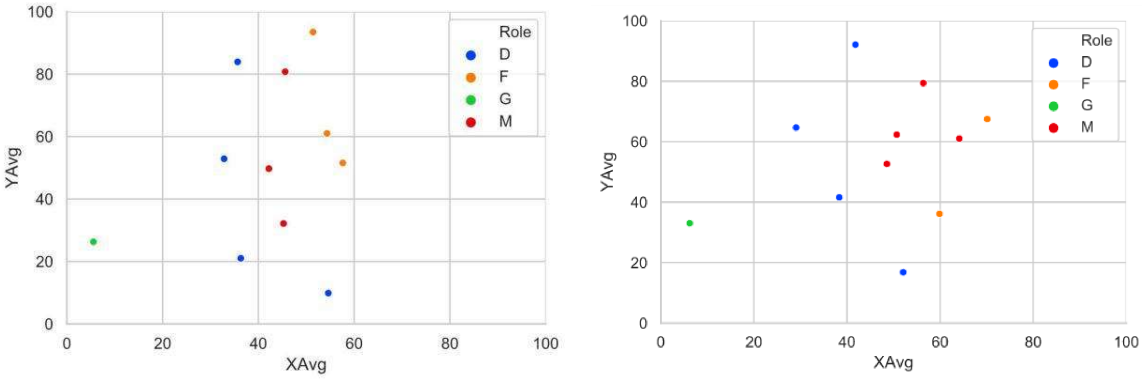

Figure 7 Average Co-ordinate (Approximate Formation)

4.2 Dynamic Index (DI)

Dynamic indicators include the team's man-made influence factors and technical data generated in the match: artificial influence factors include coaches, opponents' levels, home or away, and technical data include statistics on various events including shooting, passing, clearance etc. The original data uses a single event as a sample unit, and we classify it as dynamic data in units of one match. By observing the data stored in the new structure, we can extract some of the feature information.

4.2.1 Data Cleaning and Feature Engineering

In feature engineering, in order to reduce the dimensionality of features, we can not only use PCA to screen and remove features that have no significant impact, but also use ChiMerge feature binning method to divide EventSubTypes into four aspects: passing, defense, and fail. These aspects along with coaches, home or away, and opponent's levels is considered to be the features of a match, and use standardized, dummy variables, combined analysis and other methods to process the statistical data to quantify it:

1. Statistical data

$$Defence(G_i) = Clearance + Blocks + Interruption + Aerial\ Dual + Saves$$
$$Attack(G_i) = Shots + Dribbles + Touch + Corners + Offside$$
$$Fail(G_i) = Loss\ of\ Possession + Fouls$$
$$Oppo(G_i) = Pt\ \ (OpponentID) + \sum_{j=1}^{38} GD_j(OpponentID)$$

2. Multi-event combined analysis data

$$Possession(G_i) = \frac{1}{90min}\sum_{i=2}^{n}(t_i - t_{i+1}),(n\ is\ the\ number\ of\ Huskies'data)$$

3. One-Hot encoded dummy variable data

$$Side(G_i) = \begin{cases} 0, home \\ 1, away \end{cases} = \begin{cases} [1,0], home \\ [0,1], away \end{cases}$$
$$\begin{cases} Coach(1) = [1,0,0] \\ Coach(2) = [0,1,0] \\ Coach(3) = [0,0,1] \end{cases}$$

4.2.2 Visualization Analysis

Analyze the effect of $Side(G_i)$ on $Goal(G_i)$ and $Ratings(G_i)$:

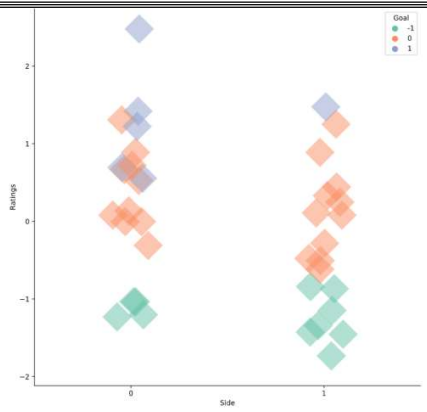

Figure 8 Goal & Ratings Function of Side

主客场与得分的关系图

When $Side(G_i) = 0$, $Goal(G_i) = 0\ or\ 1$ has more distribution, and $Ratings(G_i)$ has more distribution. Therefore, the overall performance at home is better than away.

Analyze the coaching levels of different coaches and the effectiveness of coaching for the team $Attack(G_i), Defence(G_i), Passes(G_i)\ and\ Fail(G_i)$:

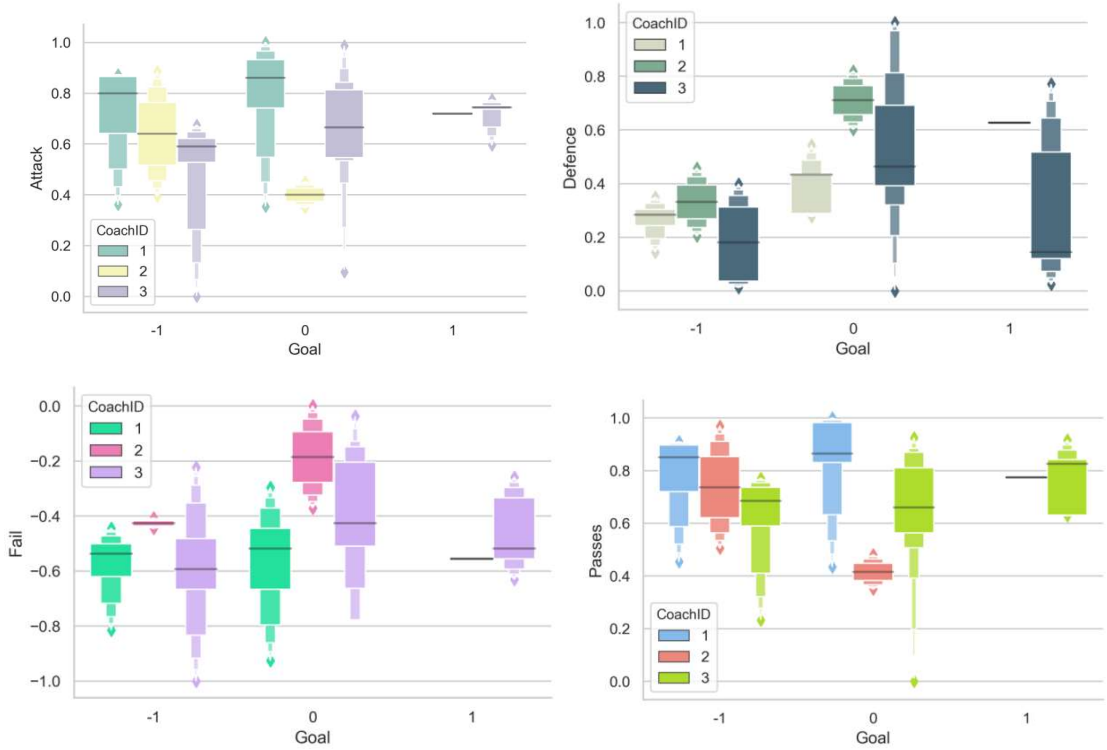

Figure 9 Performance with Different Coach

From the figures, we can find that under the guidance of Coach 3, the team's $Goal(G_i), Attack(G_i)$ and other features are better, followed by Coach 2 and Coach 1. We can also show their coaching styles, for example, coach 1 is more aggressive, while his defense is mediocre; coach 2 emphasizes violent defense; coach 3 is more balanced and has the best performance.

Analysis of $Attack(G_i)$ and $Passes(G_i)$'s contribution to $Goal(G_i)$:

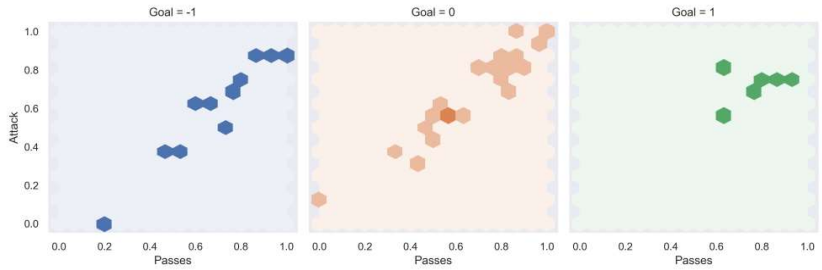

Figure 10 Attack & Passes Relation with Goal

From the figure we can find that under different goal difference numbers, the attack and the pass are generally linearly related, with a positive slope.

$Goal(G_i)$ is positively correlated with $Passes(G_i)$ and $Attack(G_i)$, and the more the distribution concentrated, the smaller the variance of $Passes(G_i)$ and $Attack(G_i)$. We can conclude that the more $Goal(G_i)$ in a match or even the entire season, the higher the probability of better $Passes(G_i)$ and $Attack(G_i)$ is.

Analysis of $Defence(G_i)$ and $Fail(G_i)$'s contribution to $Goal(G_i)$:

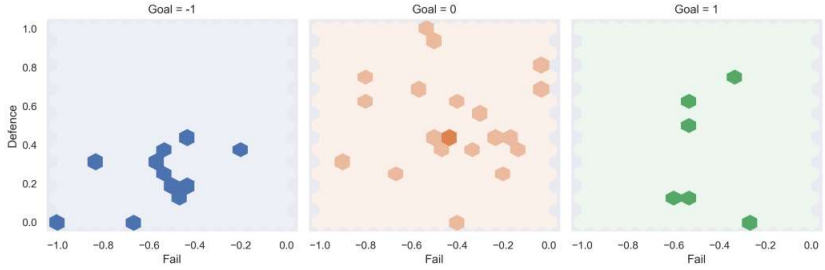

Figure 11 Defense & Fail Relation with Goal

$Goal(G_i)$ is positively correlated with $Defence(G_i)$, negatively correlated with $|Fail(G_i)|$, and the more the distribution concentrated, the smaller the variance of $Defence(G_i)$ and $Fail(G_i)$. Observations: The points on the left of Figure 12 are distributed at the bottom of square, so a bad defense will lead to a loss; there is no point on the left half of Figure 12, so you can't make too many mistakes if you want to win.

Take $Attack(G_i), Defence(G_i), Passes(G_i)$ as the positive indicators for examining the overall performance of the team, along with $Passes(G_i)$, $Oppo(G_i)$ for multi-angle analysis:

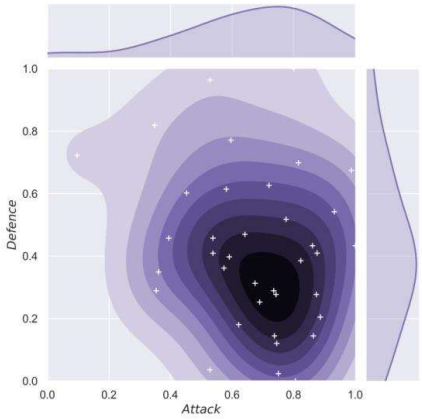

Figure 12 Oppo & Pass & Side

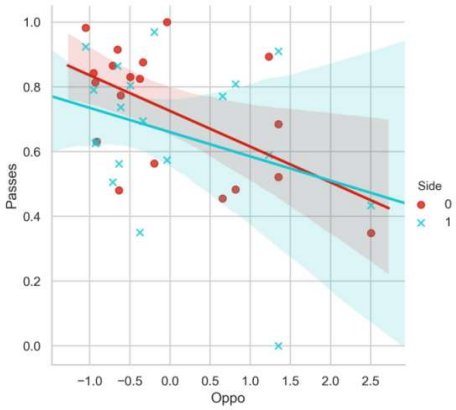

Figure 13 Attack & Defense

From the picture on the left, we can find that the center of gravity of the data is distributed in

the lower right corner. It is believed that $Attack(G_i)$ (attacking performance) is significantly better than $Defence(G_i)$ (defensive performance) throughout the season. From the picture on the right, we can find that $Passes(G_i) \propto \left[\alpha \frac{1}{Oppo(G_i)} + \beta\right]$, whether at home or away, but it is more likely to have a small improvement at home. The conclusion is that the higher the opponent's level, the lower the relative value of our pass.

Synthesize all the processed features, and estimate the correlation of the pairwise features among the variables by calculating the Pearson correlation coefficient.

$$r_{xy} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}}$$

Let the matrix $Arr[i,j] = r_{ij}$:

Figure 14 Correlation Coefficient Matrix From DI

4.2.3 RFC Establishment, Optimization, and Training

We use $Goal(G_i)$ as the evaluation label for each match. We hope that the learned model can classify the match based on the processed data and correspond to the $Goal(G_i)$ label. Due to M = 10 features are too many and their correlations with labels are different, it is not appropriate to use a linear model for classification; and the number of sample data N = 38 is very small, so it is easy to have large deviations when using some deep learning algorithms. In summary, we choose a random forest model to build a $Goal(G_i)$ label classifier.

Random forest is a classifier containing multiple decision trees, and the output is determined by the mode of the output by individual trees. For many kinds of data, it can generate a high-accuracy classifier; it can evaluate the importance of variables when determining categories; and when it builds forests, it can internally produce unbiased estimates of generalized errors. The method for establishing a Random Forest Classifier is as follows:

1. Input the number of features m, which is used to determine whether the decision result of a node on the decision tree meets $m < \sqrt[2]{M}$
2. Use Bootstrap sampling to sample N times from the N samples with a sampling method to form a training set, and use the unselected samples as predictions to evaluate their errors
3. For each node, randomly select m features. The decision of each node in the decision tree is determined based on these features. Calculate the best splitting method based on these m characteristics;
4. Each tree will grow completely without pruning. This may be used after building a normal tree classifier.

After training the random forest classifier, use grid search to optimize the parameters and select

$$\begin{cases} n\_estimator = 50 \\ random_{rate} = 0 \\ \max\ \_depth = 3 \\ \max\_feature = \sqrt[2]{M} \end{cases}$$

as parameters, the K-fold cross validation test was used to calculate its accuracy score, which was used to evaluate the accuracy of the model.

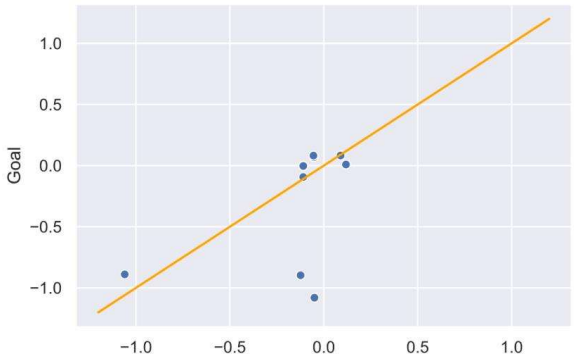

Figure 15 Result of Cross Validation

After certain data adjustments and simulation results, the average score is 65.8%, and the best situation can reach a score of 80-90%. When the sample size is only N = 38, we can accept the accuracy of this model to predict the goal difference by dynamic indicators.

5 Design of Structural Strategies Driven by SA

The structure strategy affects the successful team cooperation, as a successful coach should have better overall planning, coordination, cooperation, personnel arrangement ability. In our opinion, the specific structure strategy should be mainly reflected in the following two aspects: player position arrangement and team formation. In addition, model should also consider the tacit understanding between the players, home and away influence, coach arrangement.

5.1 Position Evaluation Engineering (PEE)

When considering the arrangement of players' positions, it is necessary to calculate the contribution value of different players in the positions of goalkeeper, striker, midfielder and defender. We collect the Eventtypes of 30 players in the Huskies in the data set, and use them as the horizontal axis and player number as the vertical axis to count the number of each Eventtype of each player in the whole season, and use the depth of color to express the number of times. The following are the Eventtypes statistics of forwards, midfielders and defenders:

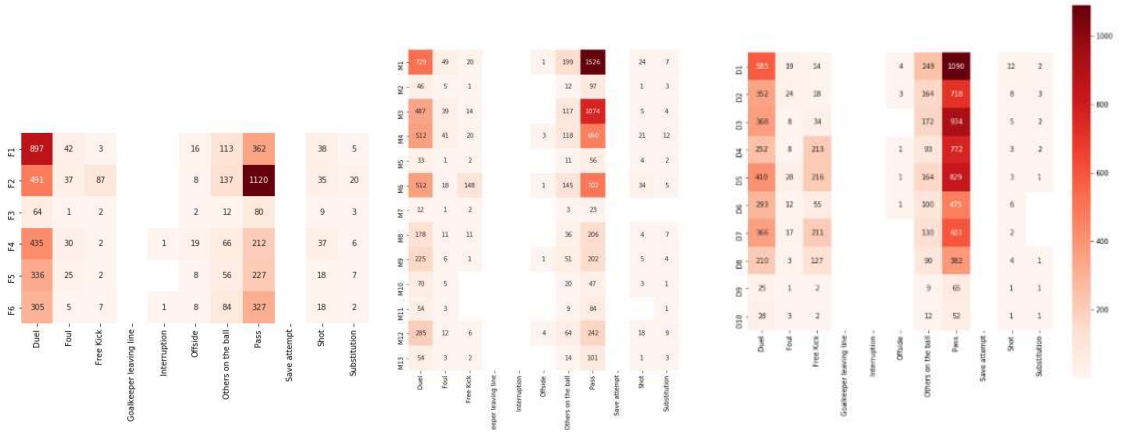

Figure 16 Forwards & Midfielder & Defender EventTypes Statistics

From the above four figures, we can see that the largest contribution of F is F2, followed by F1, F6, F5, F4. In M, M1 is the largest contributor, followed by M3, M4 and M6. In D, the largest contribution is D1, followed by D3, D5, D4, D2, D7, D6, D8.

We hope to have a practical model to evaluate the performance of different players in different positions. At this time, it is necessary to analyze the important data of different positions in combination with practical knowledge, calculate the weight distribution of different event types, and combine the performance of various abilities of players. The performance of 29 players (except the only goalkeepers) as evaluation of the team in G, F, M positions respectively. In the following figure, the redder the color is, the more suitable the position is; otherwise, the bluer it is, the less suitable it is.

Figure 17 Evaluation varies from players and positions

5.2 Optimization of permutation and combination based on SA algorithm

We analyze the line-up of the main line-up in 38 matches of the whole season, and hope to build a model to suggest the best team lineup for the coach. The goal of this model is to find an optimal orderly combination, so that the sum of the abilities of 11 players in their respective positions is the largest. The 11 positions on the field are arranged in order, and the current status is represented by the 30 HEX gray code; for example, the gray code '0a1grd739ki' indicates that there are players 0, 10, 1, 16, 26, 13, 7, 3, 9, 11 and 18 in turn. In the case of huge search tree and limited computing resources, we choose simulated annealing algorithm. One of the main advantages of the simulated annealing algorithm is that it can accept the state with a certain probability that the value of the objective function is not good, and it can continuously accept the solution that makes the objective function move in a good direction in the process of iteration. The specific steps of simulated annealing algorithm are as follows:

1. Give the parameters of cooling schedule and initial solution of iteration x_0, and $f(x_0)$, the parameters of cooling schedule include: Initial value T_0 of control parameter T, attenuation function, final value and chain length of control parameters L_k;
2. When the parameter $T = T(k)$, perform L_k exploratory searches as follows:
   a) According to the properties of the current solution X_k, a random offset m is generated, and a new trial point X_k of the neighborhood of the current solution is obtained;

b) Generate a random number θ uniformly distributed on the interval $(0,1)$, and calculate the transfer probability P corresponding to the acceptance criteria given the current iteration point X_k and temperature T_k:

$$P = \begin{cases} 1, & f(X_k') < f(X_k) \\ \exp\left(\dfrac{f(X_k) - f(X_k')}{T_k}\right), & f(X_k') < f(X_k) \end{cases}$$

$$Attitude(X_k) = \begin{cases} Accep\ , \theta < P \\ Reject, \theta \geq P \end{cases}$$

c) If the exploratory search is less than L_k times, return to step 1, otherwise go to step 3;

3. According to the given temperature decay function, a new temperature control parameter T_{k+1} and chain length L_{k+1} are generated. Turn to step 2, and enter the optimization of the equilibrium point of the next temperature point.

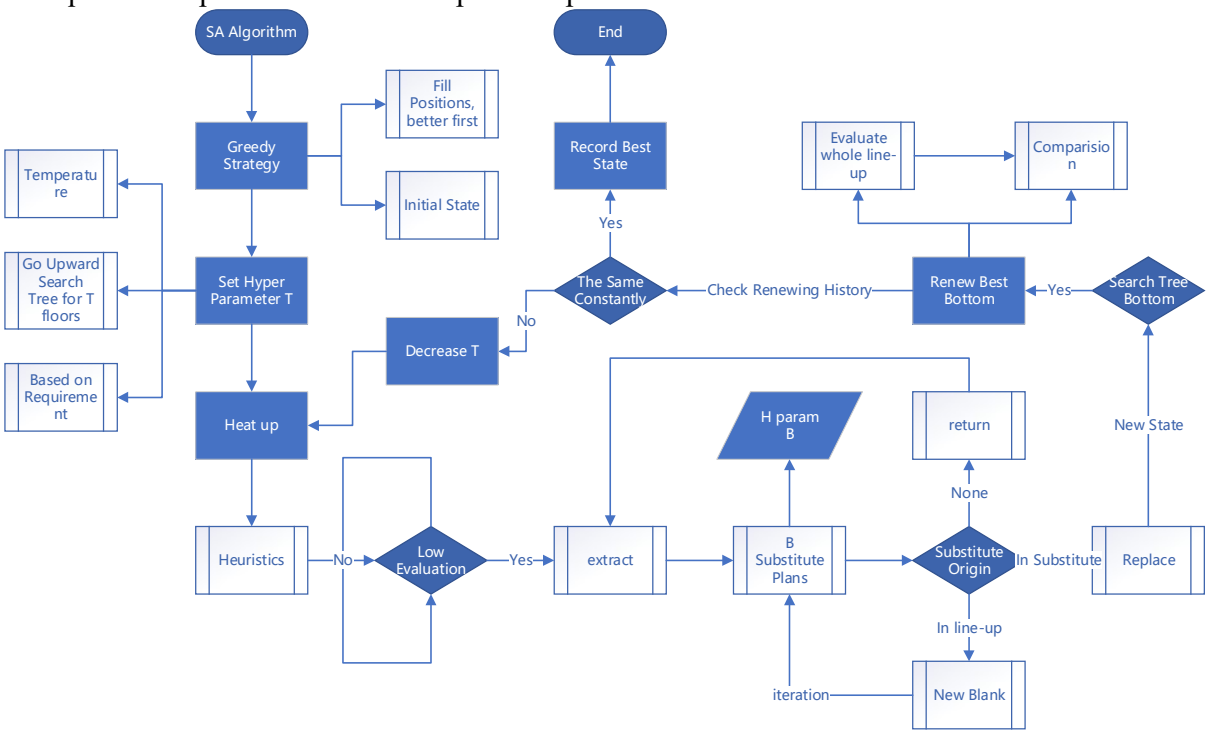

Figure 18 SA Flowchart

In the actual exploratory search, we are likely to enter the local optimum, and need to make a decision to exit. When the optimization degree of the current solution is less than that of the current optimal solution, the probability of the new solution being accepted is, while when the temperature is low enough, the probability of the worse solution being accepted tends to be. According to the feature that there is no more optimized solution in the recent sub search, the elucidation value can be determined according to the specific problem and then it can be determined that the search has entered the local optimum.

5.3 Other Structural Strategy Factors

After considering the main strategy, we considered the following four secondary factors: players' rapport, home and away influence, and coaching arrangements.

First of all, choosing a team with a high degree of tacit understanding is conducive to improving the efficiency of passing and scoring. Teams with a high degree of tacit understanding

often have a strong ability to cooperate, which contributes to the success of the matches. Players with higher passing efficiency tend to be more adaptable and cooperate better with other players.

Home and away factors must also be considered, some players are more adaptable, at home and away can better play the original level, while some players are less adaptable only at home to play the original level, the environment has a greater impact on his performance. Then at home and away, different players should play.

Finally, in terms of Coach arrangement, Coach 1, Coach 2 and Coach 3 respectively guided 9,5 and 24 matches in the whole season. According to the data analysis in the second question, it can also be concluded that Coach 3 has a higher level.

5.4 Structural Strategy Conclusion

Throughout the entire model, in order to improve the win rate next season, our advice is to hire Coach 3, the team Coach, use the 442 line-up, make F1, F2, F6, M3, M1, M6, D1, D2, D3, D5 and G1 as main force, naming the formation as $lineup_0$, Their positions are arranged according to the figure below：

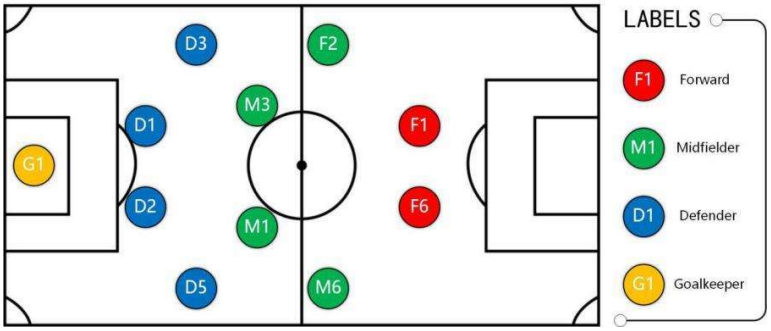

Figure 19 Best Line-up

According to the evaluation of the season data, F2 player, as a striker, has a strong ability in the midfield. When trying to arrange him in the midfield, a significant new optimal solution is obtained, which indicates that the evaluation of any position of everyone is more important.

To sum up, the sum score of personal ability of the formation is $PersonalScore(lineup_0) = 94.43$, the score of team cooperation is $CoordinationScor\ (lin\ \ _0) = 90.17$, and the weighted average is based on $\begin{cases} 0.7, Personal \\ 0.3, Coordination \end{cases}$, and the final comprehensive score is $TotalScore(lineup_0) = 93.152$. In the actual competition, the formation similar to this has achieved good results, and also verified the feasibility and accuracy of our evaluate model and simulated annealing algorithm.

6 Model Extension Combined with Group Dynamics

In the research of the Huskies, we found some factors that affect the successful team cooperation, such as passing network, personal ability, coach and so on. These factors can be linked to group dynamics to analyze why these football field factors contribute to team performance. And we can explore which factors can be considered to supplement our interpretation of excellent groups and spread them into various groups in society.

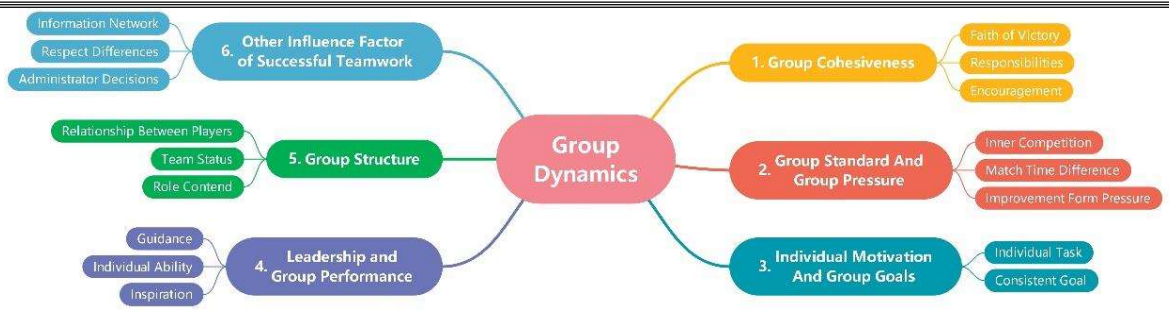

Figure 20 Group Dynamics Mind Map

6.1 Group and Soccer Team

Group dynamics mainly includes 5 aspects: group cohesiveness, group pressure and group standards, individual motivation and group goals, leadership and group performance and group structure.

6.1.1 Group Cohesiveness

Group cohesion can be regarded as the belief that the team wants to win. Group pressure comes from the outside. Improving group cohesion is a virtuous circle to inspire the team's winning belief. Responsible behavior is reflected in each player's performance. The mutual influence among the members is reflected in the mutual encouragement and progress among the players.

The Huskies is under the pressure of their opponents in the competition, but also full of victorious spirit to win. In data analysis of Huskies, each players' average acceleration, running distance and effective area of heatmap can reflect the particular player's attitude and effort on the court. In the situation of backwardness, draw and lead, the team bears varying degrees of external pressure, as the result, the data of players' attitude will fluctuate. If there is still positive data in the situation of draw and backwardness, we can consider that the group cohesion is really strong.

6.1.2 Group Standard and Group Pressure

Group standards can help teams and players to feel more oppressive and to compete with each other, which helps improve personal ability. From this point of view, appropriate group pressure is necessary, and only when there is competition can there be development.

For Huskies, there are differences in minutes. There is a huge gap between the core players and the edge players, which leads that the edge players in the team will be under the pressure of the core players' ability and status, but it also encourages them to strive to get playing time to prove themselves. We can analyze the evaluation trend of every match of players, especially the players who can't get stable playing time. If their evaluation can be improved in limited playing chances, we can think that the pressure brought by internal competition makes them progress in a way.

6.1.3 Individual Motivation and Group Goals

Group goals affect group behavior. When team goals and players' goals are consistent, players will show the strongest motivation to win and strive for goals.

Players in every position of the team have different responsibilities, and the victory is that all players have completed their tasks perfectly. Therefore, everyone is the same in the goal of victory, and everyone has certain positive expectations for their own tasks. When everyone works hard for the expectation of completing their own tasks at the same time, the group goals will reach an agreement.

6.1.4 Leadership and Group Performance

There are two kinds of leaders in the team, they are the coaches who provide training, pre-match guidance and post-match summary for the matches, or the leaders who inspire morale, set an example and dispatch the command on the field. The ability of leaders will affect the team's performance and progress. In addition, the coach and the captain's help to the players can improve the team's vitality.

6.1.5 Group Structure

When a team has a stable relationship between players, it has a team structure. Stable structure is conducive to the cultivation of tacit understanding among players, so as to have a greater probability of winning the match. It can be convinced that the number of people in different positions of a team obeys the upper triangle distribution, and if the position range is too large or too small, it is not conducive to the stability of the team. In addition, if the competition in one position of the line-up is too great and the competition in other positions is very small, there will be structural imbalance and structural changes.

6.2 Other influence factor of successful teamwork

There is also a need for close and unblocked information exchange network between teams. International teams need to ensure that the language communication between players can be smooth on and off the court; in addition, they need to carry out more friendship activities to improve the harmonious atmosphere within the team. In this way, we can communicate and cooperate effectively in the competition.

Leaders and everyone should respect the differences between others and themselves, and improve the team atmosphere by accepting the differences between people while maintaining group goals and cohesion.

Administrators should have a clear understanding of the situation and be able to make adjustments to the situations.

7 Evaluation

7.1 Strength

1. The establishment of PNM is closely related to the design of PEI. PEI comprehensively considers many aspects of each pass, quantifies the quality of the pass, and can reduce the error and variance with the actual situation. And the network model of pass, based on graph theory, intuitively describes the degree of cooperation, which is conducive to the search of multiple combinations, and the visual effect can highlight the familiar combination.
2. Heatmap generation model has strong compatibility for the approximate continuity of discrete data, and can cope with the situation of too little or sparse coordinate data. Based on the visual data analysis, player position is consistent with the actual situation. In addition, for the cleaning of events data, the impact of data abnormality and missing is effectively avoided. Even though the number of samples is only Under 38, it is not easy to overfit or deviate too much. The highest accuracy of 80% after parameter optimization is enough to effectively predict the general results of the match, which means this model can make predictions for the future match, that reflects the current ability of the team, based on the recent data and give training and line up to coaches as reference.
3. The static structure strategy, which should be developed by the coach, is transformed into the

optimal arrangement and combination problem of 11 elements. Large-scale data supports the increase of the dimension of the evaluation index and reduces the expectation of the deviation of the particular position ability value of each player. Under the condition of limited computing power, the simulated annealing algorithm is properly used, and we manually set the starting strategy according to the actual experience and uses the imprecise individual ability evaluation index to find the arrangement and combination of 11 players. The partial optimal solution can be accepted as the global optimal solution within the threshold range when the accuracy expectation is certain.

4. The models can be easily corresponded to the theoretical key points in group dynamics, and the additional aspects based on the existing influencing factors are also of great practical significance.

7.2 Weakness

- There are many hyper parameters in the model, so the parameter optimization of the model has a great challenge.
- There are few players with data at each time, so it is impossible to test the players' ability without the ball, and it is difficult to evaluate the attack or defense from the aspect of the overall position and formation.
- There are too few samples to input into the RFC model, so the training results fluctuate greatly.
- The optimal lineup obtained by the simulated annealing algorithm can only be guaranteed to be a local optimal solution, not a global optimal solution.

8 Reference

[1] Variable selection using random forests[J] . Robin Genuer,Jean-Michel Poggi,Christine Tuleau-Malot. Pattern Recognition Letters . 2010 (14)

[2] Random Forests[J] . Leo Breiman. Machine Learning . 2001 (1)

[3]Multi-objective optimizationof simulated countercurrent moving bed chromatographic reactor foroxidative coupling of methane. KUNDUP K,ZHANG yan,RAY A K. Chemical Engineering Science . 2009

[4]An effective hybrid particle swarm optimization for no-wait flow shop scheduling[J] . Bo Liu,Ling Wang,Yi-Hui Jin. The International Journal of Advanced Manufacturing Technology . 2007 (9-10)

[5]Hu Bo, Li Chao, Huang Fu. Calculation of indirect adjustment based on graph theory model [J]. Urban survey, 2019 (06): 130-132

[6]Liu Sihan, Shang Xiaming, Ma ting. Classification of forest types based on random forest feat ure selection [J]. Beijing mapping, 2019,33 (12): 1518-1522

[7]Technology - Information Technology; Reports Outline Information Technology Study Result s from Army Engineering University (Cloud Annealing: A Novel Simulated Annealing Algorith m Based on Cloud Model)[J]. Computers, Networks & Communications,2020.

[8]Xiao Xiaowei, Xiao Di, Lin Jinguo, Xiao Yufeng. Overview on multi-objective optimization p roblem research [J]. Application Research of Computers, 2011,28 (03): 805-808+827.

[9]Chu Tianguang, Yang Zhengdong, Deng Kuiying, Wang Long, Xie Guangming. Problems in s warm dynamics and coordinated control [J]. Control Theory & Applications, 2010,27 (01): 86-9 3.