NBA における新ポジションの提案とラインナップ最適化

Proposing new positions and optimizing player combinations in the NBA

令和 3 年度 修士論文
横浜市立大学大学院
生命ナノシステム科学研究科
物質システム科学専攻
塚田 憲哉

目 次

NBA における新ポジションの提案とラインアップ最適化
Proposing new positions and optimizing player combinations in the NBA

物質システム科学専攻 塚田 憲哉
指導教員 Micheletto Ruggero

ソフトクラスタリング：クラスタに所属する確率を計算し，適切なクラスタ数と，各クラスタに所属する確率であるクラスタ確率を算出する．
ラインナップ：バスケットボールにおけるコートに同時に出場する 5 人の選手の組み合わせ．
OffRTG：100 回の攻撃における得点数．
DefRTG：100 回の守備における失点数．
ラインナップ最適化：説明変数をクラスタ確率，目的変数を OffRTG．DefRTG として決定木ベースの回帰分析を行い，最適なラインナップを予測する
NBA：北米男子プロバスケットボールリーグ

＜研究の背景と目的＞

バスケットボール競技において，試合中に 1 度にプレーすることができるのは 5 人のみである．また，選手はそれぞれ 5 つのポジションが割り当てられている．しかし，近年の NBA では，選手の役割の多様化によりポジションレス化が進んでおり，多くの選手の役割や特徴を従来のポジションに当てはめることが難しい．

先行研究において，NBA データを用いたクラスタリングにより，選手の特徴を正確に捉えることができる新たなポジションが提案された．新ポジションの提案には，教師なしクラスタリングである Gaussian Mixture Model (GMM) を使用し， 9 つのポジションを提案した．その新ポジションから，ランダムフォレストを用いて選手 5 人の組み合わせであるラインナップ最適化を行った．

先行研究の課題として，主にオフェンスのスタッツを使用しており，ディフェンスの能力が十分に評価されていない．また，オフェンスで同様の特徴を持っている選手も，ディフェンスでは異なる特徴を持つ場合がある．さらに，ラインアップを構築する上で選手のスキルが考慮されていないという課題もある．

本論文では，試合中の細かなプレーに関するスタッツとディフェンスに関するスタッツを追加し，オフェンスとディフェンスそれぞれの新ポジションの提案を目指す．また，提案された新ポジションと選手のスキルを組み合わせることで，ラインナップ最適化も行う．研究結果から，チーム戦略や移籍の支援だけでなく，組み合わせ最適化モデルを用いた他分野への応用も目指す．

＜方法＞

対象とするデータは，NBA の 2015-16 シーズンから 2020-21 シーズンのレギュラーシー

ズンにおける全チームの試合データとする．なお，選手個人のデータは異常値を除去する為にシーズン通算 30 試合以上出場とし，1 試合平均 12 分以上出場した選手とした．クラスタリングでは選手の能力よりもプレーの特徴を捉える為に，シュートに関するスタッツの多くは成功率ではなく試投数とした．また，出場時間によるパフォーマンスの差を除去する為に，シュート試投数は全フィールドゴール試投数に対する割合とし，その他スタッツは出場時間に対する割合とした．クラスタリングには統計解析ソフト R の mclust パッケージによる EM アルゴリズムに基づくソフトクラスタリングを用いた．mclust を使用することで．クラスタに所属する確率を得ることができる．

ラインアップ分析では，クラスタリングによって割り当てられたクラスタを新ポジションとし，各クラスタに所属する確率を用いて回帰分析を行う．表 1 のようにクラスタ確率を基にラインアップを作成し，各ポジションの合計値を説明変数とする．目的変数はラインアップの評価指標である Offensive Rating，Defensive Rating を用いた．各 Rating は 100 ポゼッションごとの得失点数を表している．

表 1: クラスタ確率を用いたラインナップの例

選手名	Cluster 1	Cluster 2	Cluster 3	···	Cluster n
選手 1	0.30	0.70	0.00	···	0.00
選手 2	0.00	1.00	0.00	···	0.00
選手 3	0.10	0.00	1.00	···	0.00
選手 4	1.00	0.00	0.00	···	0.00
選手 5	0.00	0.00	0.00	···	1.00
合計値	1.40	1.70	1.00	···	1.00

クラスタ確率を使用した回帰分析を行った後，選手のスキルとクラスタ確率の両方を考慮したラインナップ最適化も行う．選手のスキルを選手版 Offensive Rating，Defensive Rating とし，それぞれにクラスタ確率を乗じたものを説明変数とする．

＜結果と考察＞

クラスタリングの結果，オフェンスは 7 つ，ディフェンスは 5 つのクラスタとなった．それぞれの分布を図 1，2 に示す．

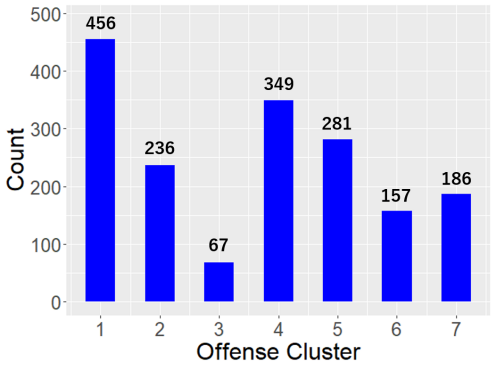

図 1: オフェンスクラスタの分布

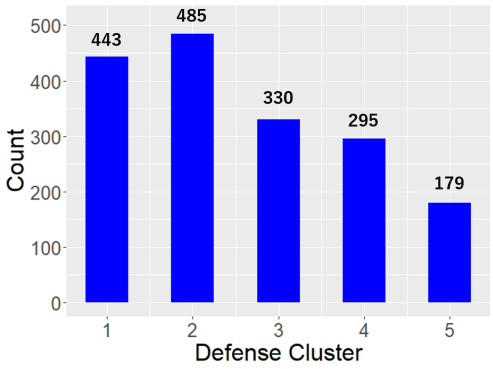

図 2: ディフェンスクラスタの分布

各クラスタのスタッツを比較し，特徴の一部を表 2, 3 にまとめ，新ポジションを定義

した.

表 2: 新ポジションの定義 (オフェンス)

新ポジション	特性	ハイスコア スタッツ	ロースコア スタッツ
Utility Forward	アウトサイドシュートの成功率が 高く，インサイドでも得点できる 汎用性の高いフォワード.	MID FGA% USG, 3P%	CAT 2P，RES OF REB
Scoring Guard	Solid Guard よりも USG が高く，シュート本数や， アイソレーションが多い.	FGA，USG AST, ISO	OF REB C3 FGA% CAT 3PFGA

表 3: 新ポジションの定義 (ディフェンス)

新ポジション	特性	ハイスコア スタッツ	ロースコア スタッツ
Rim Protector	ブロック数が最も多く， リング周辺でディフェンス する為移動速度は遅い.	DF REB，BLK CON 2P	STL，DFL CON 3P
Versatile Defender	スティール以外のスタッツは 平均的に高く，汎用性が高い.	dfISO，DF REB CON 3P	STL，DFL

オフェンスとディフェンスに分けてクラスタリングを行うことで，従来のポジションに比べ，より選手の特徴を捉えた新ポジションを提案することができた．また，表 4 のような予測用データセットを使用し，スキルを考慮したラインナップ最適化モデルを作成した．各ポジションの合計値が 450 から 600 までの全ての組み合わせを，間隔を 30 として作成し，予測モデルから予測値を求めた.

作成したモデルを用いて，実際に Los Angels Lakers の Lebron James と Anthony Davis の 2 選手を中心に Offensive Rating が最大となるようなラインアップを構築した．その結果，構築したラインアップの方が，実際の Los Angels Lakers のラインナップデータの最高値よりも高い値となった.

表 4: 予測用データセット (クラスタ確率 ×Player OFF RTG)

Solid Guard	Utility Forward	High Usage Forward	Spot Up Shooter	Stretch Rebounder	Traditional Center	Scoring Guard
600	0.00	0.00	0.00	0.00	0.00	0.00
570	30	0.00	0.00	0.00	0.00	0.00
540	30	30	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...
0.00	0.00	0.00	0.00	0.00	0.00	600

オフェンスにおいては，選手のスキルをクラスタ確率と共に使用することで，実際の選手のスキルを考慮した予測モデルを構築する事ができた．これにより，選手の移籍やチームの戦術の提案にも繋がる結果が得られた．しかし，ディフェンスにおいては予測モデルが極端な値をとってしまい，ポジション間での相互作用を正確に示すことができなかった．ディフェンスのスキルとクラスタ確率を使用した予測モデル構築では，クラスタリングに

使用したスタッツが 11 個であり，オフェンスに比べ少なかったため，十分にディフェンススキルを評価できなかったと考えられる．そこで，より詳細なディフェンスのスタッツを追加することで精度の向上に繋がると考えられる．本論文では，オフェンスとディフェンスに分けてポジションを提案したが，今後の展望として，それらを組み合わせて 1 つの予測モデルを作成することができれば，より実戦で使いやすくすることができると考えられる．

1　はじめに

バスケットボール競技において，試合中に 1 度にプレーすることができるのは 5 人のみである．また，選手はそれぞれ 5 つのポジションが割り当てられている (NBA.com, 2015)．しかし，バスケットボールは近年，選手の役割の多様化によりポジションレス化が進んでおり，多くの選手の役割や特徴は，表 1 に示す従来の 5 つのポジションに当てはめる事が難しい．

表 1: 従来の 5 つのポジション

略称	名前	役割
PG	Point Guard	チームの最高のドリブラーでありパサーである． 相手の PG を守りボールのスティールを狙う．
SG	Shooting Guard	チームの最高のシューターである． 長距離のシュートを決めることができ．良いドリブラーである．
SF	Small Forward	背の低い選手や高い選手を相手にプレーする．コート上を 動き回り，長距離でも近距離のシュートでも得点することができる．
PF	Power Forward	リングの近くでリバウンドや，背の高い選手をディフェンス するなど，センターのような役割を担う． しかし，センターよりも長距離のシュートを打つことがある．
C	Center	チームで最も背の高い選手で，リングの近くでプレーする． オフェンスでは近距離のシュートでの得点やリバウンドをとる． ディフェンスではシュートブロックや，シュートのリバウンドをとる．

先行研究において，北米男子プロバスケットボールリーグ National Basketball Association (以下：NBA) のデータを用いたクラスタリングにより，現代の選手の特徴を正確に捉える事ができる新ポジションが提案された (Kalman and Bosh, 2020)．新ポジションの提案には，教師なし学習のクラスタリングである Gaussian Mixture Model (GMM) を使用し，9 つのポジションを提案した．その新ポジションから，選手 5 人の組み合わせであるラインアップの最適化を，ランダムフォレストを使用することで行った．

先行研究にて使用されたスタッツの多くはオフェンスに関するものであり，ディフェンスの能力は十分に考慮されていない．また，オフェンスにおいて同様の特徴を持つ選手であっても，ディフェンスでは異なる特徴を持つ場合がある．さらに，ラインアップを構築する上で，選手のスキルが考慮されていないという課題もある．そこで，本研究では試合中の細かなプレーに関するスタッツと，ディフェンスに関するスタッツを追加し，オフェンスとディフェンスに分割することでそれぞれの新ポジションの提案を目指す．また，提案された新ポジションと選手のスキルを組み合わせることでラインアップ構築も行う．研究結果から，チーム戦略や選手の移籍の援助だけでなく，組み合わせ最適化モデルを用いた他分野への応用も目指す．

2　本論文で用いるスタッツ

本論文でクラスタリングに用いるスタッツを以下の 表 2，3 に示す．オフェンスではシュートに関する基本的なスタッツに加え，ポストアップ数やキャッチアンドシュート数，

エリア別のシュート数等も使用することで，スペーシングや選手の細かな特徴を考慮してクラスタリングを行う．USG の計算方法を式 1 に示す．なお，Possession とは攻撃回数を表す．

表 2: オフェンスのクラスタリングに用いるスタッツ及び記号

記号	スタッツ	意味
HT	Height	身長
PTS	Points	得点数
FGA	Field Goals Attempted	フィールドゴール試投数
3PA%	3 Point FGA / FGA	フィールドゴール試投数に対する 3 ポイントシュート試投数の割合
3P%	3 Point Field Goal %	スリーポイント成功率
FTA	Free Throws Attempted / FGA	フィールドゴール試投数に対する フリースローの割合
FT%	Free Throw %	フリースロー成功率
AST	Assist Adjusted	アシスト数にフリースローと セカンダリーアシスト数を加えた総数
OF REB	Offensive Rebounds	オフェンスリバウンド数
TOV	Turnovers	ターンオーバー数
DRI	Drives	ドライブ数
USG	Usage Rate	使用率 (式 (1) 参照)
ISO	Isolation	アイソレーション数
PUL FGA%	Pull Up FGA / FGA	フィールドゴール試投数に対する プルアップシュート試投数の割合
POS	Post Ups	ポストアップ数
SPO FGA%	Spot Up FGA / FGA	フィールドゴール試投数に対する スポットアップシュート試投数の割合
CAT 2PFGA%	Catch & Shoot 2 Point FGA / FGA	フィールドゴール試投数に対するキャッチアンド シュート 2 ポイントシュート試投数の割合
CAT 3PFGA%	Catch & Shoot 3 Point FGA / FGA	フィールドゴール試投数に対するキャッチアンド シュート 3 ポイントシュート試投数の割合
TRN	Transitions	トランジション数
RES FGA%	Restricted Area FGA / FGA	フィールドゴール試投数に対する 制限区域内シュート試投数の割合
ITP FGA%	In The Paint FGA / FGA	フィールドゴール試投数に対するペイントエリア内 シュート試投数の割合
MID FGA%	Mid-Range FGA / FGA	フィールドゴール試投数に対するミドルレンジ シュート試投数の割合
C3 FGA%	Corner 3 Point FGA / 3PA	3 ポイントシュート試投数に対するコーナー 3 ポイントシュート試投数の割合
TOUCH	Touches	タッチ数
PASS	Passes	パス数

表 3: ディフェンスのクラスタリングに用いるスタッツ及び記号

記号	スタッツ	意味
HT	Height	身長
DF REB	Defensive Rebounds	ディフェンスリバウンド数
STL	Steals	スティール数
DFL	Deflections	ディフレクション数
BLK	Blocks	ブロック数
DFG%	Dfended Field Goal %	ディフェンスした相手のフィールドゴール成功率
CON 2P	Contested 2 Ponit Shots	シュートチェックした 2 ポイントシュート試投数
CON 3P	Contested 3 Point Shots	シュートチェックした 3 ポイントシュート試投数
dfISO	Defended Isolations	アイソレーションに対してのディフェンス回数
dfPOS	Defended Post Ups	ポストアップに対してのディフェンス回数
SPE	Average Speed Defense	ディフェンス時の平均移動速度

$$\frac{FGA + PossessionEndingFTA + TOV}{Possessions} \tag{1}$$

チームとラインアップ，選手個人に対する評価指標を以下の表 4 に示す．

表 4: 決定木分析に用いるスタッツ及び記号

記号	スタッツ	式	意味
OffRTG	Offensive Rating	$100 \times \left(\frac{Points}{Possessions}\right)$	100 ポゼッションごとの得点数
DefRTG	Defensive Rating	$100 \times \left(\frac{Points}{Possessions}\right)$	100 ポゼッションごとの失点数

Off RTG と Def RTG は 100 ポゼッションごとの得点数と失点数であり，ラインアップと選手個人の評価指標として採用する．また，選手個人の評価指標に使用する際には，該当選手がコートにいる場合の 100 ポゼッションにおける得失点数とし，それぞれ Player OffRTG，Player DefRTG とする．

3 手法

3.1 対象データ

NBA の 2015-16 シーズンから 2020-21 シーズンのレギュラーシーズンにおける全チームのデータを対象とした．なお，対象とする選手個人のデータはシーズン通算 30 試合以上出場とし，1 試合平均 12 分以上出場した選手とする．これは，30 試合以下または 12 分以下の出場時間では，各スタッツが極端な値をとることがあり，それを除外する為である．ラインアップデータは 2015-16 シーズンから 2020-21 シーズンにおける 5 シーズン全てのデータを対象とした．なお，各スタッツはシーズン平均値とし，データは NBA.com にて公開されているデータを使用した．

3.2 クラスタリング

クラスタリングの目的として，オフェンスとディフェンスのそれぞれの特徴をより捉えた新ポジションを提案したい．しかし，同じオフェンスの特徴を持つ選手同士でも，ディフェンスにおいては異なる場合がある．そこで，オフェンスとディフェンスのそれぞれでクラスタリングし，新ポジションを提案する．

クラスタリングでは選手の能力よりもプレーの特徴を捉える為に，シュートに関するスタッツの多くは成功率ではなく試投数とした．また，出場時間によるパフォーマンスの差を排除するために，シュート試投数は全フィールドゴール試投数に対する割合とし，その他スタッツは出場時間 (分) に対する割合とした．クラスタリングを行う前処理として，全てのスタッツを標準化し，平均を 0，分散を 1 とした．

対象とした試合における表 2，3 のスタッツを用いて，EM アルゴリズムによる選手のクラスタリングを行う．EM アルゴリズムを用いたソフトクラスタリングを行うことで，クラスタに所属する確率と，最適なクラスタ数を得ることができる．EM アルゴリズムについては主に松井・小泉 (2019) を参照した．

3.2.1 混合分布モデル

観測されたデータ x に確率分布モデルを想定し，どのような確率分布モデルに従って生成されたかを考える場合，正規分布が用いられることが多い．しかし，複雑なデータ発生の確率構造は捉えられない．そこで，いくつかの正規分布を線形結合で合わせた式 (2) に従う確率分布モデルを用いる．

$$f(x|\boldsymbol{\theta}) = \sum_{j=1}^{g} \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(x-\mu_j)^2}{2\sigma_j^2} \right\} \tag{2}$$

ただし，$\pi_j (j = 1, 2, \cdots, g)$ は $0 \leq \pi_j \leq 1$ を満たし，その和は $\sum_{j=1}^{g} \pi_j = 1$ とする．この確率分布モデルのパラメータは，$\{\mu_j, \sigma_j^2; j = 1, 2, \cdots, g\}$ に加えて $\{\pi_j; j = 1, 2, \cdots, g\}$ であり，これらをパラメータベクトル $\boldsymbol{\theta}$ で表す．モデルのパラメータをデータから推定し，さらに重ね合わせた正規分布の個数 g を選択できれば，いくつのクラスタに分類されるかがわかる．このように，いくつかの確率分布を線形結合で重ね合わせた確率分布モデルは，混合分布と呼ばれ次のように定式化することができる．

いま．p 次元データ x は， g 個の密度関数 $f_j(\boldsymbol{x}|\boldsymbol{\theta})(j = 1, 2, \cdots, g)$ の線形結合で表される式 (3) の確率分布に従って観測されたとする．

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{g} \pi_j f_j(\boldsymbol{x}|\boldsymbol{\theta}_j) \tag{3}$$

ここで，$\pi_1, \pi_2, \cdots, \pi_g$ は，$0 \leq \pi_j \leq 1$ を満たす混合比率で，その和は $\sum_{j=1}^{g} \pi_j = 1$ とする．分布を既定するパラメータは $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_g, \pi_1, \pi_2, \cdots, \pi_g\}$ である．特に，平均ベクトル $\boldsymbol{\mu}_j$, 分散共分散行列 $\sum_j$ の多次元正規分布 $N_p(\boldsymbol{\mu}_j, \sum_j)$ の線形結合で表される確率分布モデルは，混合正規モデルと呼ばれる．

混合分布モデルによるデータの分類は，ベイズ理論の枠組みで捉えると，次のように述べることができる．確率分布 $f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)$ を第 j 番目の群 G_j を特徴づける分布と考える．これは，データ $\boldsymbol{x}$ が第 j 群からのものであるとしたときの条件付き尤度である．また，対応する混合比率 π_j とは，その群を選択する確率，すなわち事前確率である．したがって，データ $\boldsymbol{x}$ が群 G_j からのものであるという事後確率は，ベイズの定理より式 (4) で与えられる．

$$p(j|\boldsymbol{x}) = \frac{\pi_j f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)}{\sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)} - \frac{\pi_j f_j(x|\boldsymbol{\theta}_j)}{f(x|\boldsymbol{\theta})}, \quad j = 1, 2, \cdots, g \tag{4}$$

データ $\boldsymbol{x}$ は，混合分布モデルが推定できれば，g 個の事後確率の中でその値が最大の群に属するとする．このように，分類の対象とするデータでモデルを推定して，事後確率の最大化によって各データを分類すれば，クラスタリング手法として用いることができる．

3.2.2 モデルの推定 EM アルゴリズム

混合分布モデルのパラメータの最尤推定値は，解析的に式として表すことはできない．このため推定値を求めるには数値的最適化法の適用が必要であるが，特に，EM アルゴリズムによる推定法が広く用いられている．EM アルゴリズムを適用すると，混合正規分布モデルのパラメータ $\{(\pi_j, \boldsymbol{\mu}_j, \sum_j); j = 1, 2, \cdots, g\}$ は次のステップを通して推定される．

混合比率，平均ベクトルと分散共分散行列 $\{(\pi_j, \boldsymbol{\mu}_j, \sum_j); j = 1, 2, \cdots, g\}$ に対して，初期値を設定する．第 t ステップ目の更新値 $\{(\pi_j^{(t)}, \boldsymbol{\mu}_j^{(t)}, \sum_j^{(t)}); j = 1, 2, \cdots, g\}$ から第 $t+1$ ステップ目は，次の E ステップと M ステップの繰り返しによって更新する．

E ステップ

第 i 番目のデータ $\boldsymbol{x}_i$ が群 G_j からのものであるという事後確率を式 (5) のように計算する．

$$p^{(t+1)}(j|\boldsymbol{x}_i) = \frac{\pi_j^{(t)} f_j(\boldsymbol{x_i}|\boldsymbol{\theta}_j^{(t)})}{\sum_{k=1}^{g} \pi_k^{(t)} f_k(\boldsymbol{x_i}|\boldsymbol{\theta}_k^{(t)})}, \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, g \tag{5}$$

M ステップ

第 $t+1$ ステップ目の混合比率，平均ベクトル，分散共分散行列を，式 (6) によって更新する．

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p^{(t+1)}(j|\boldsymbol{x}_i), \quad \boldsymbol{\mu}_j^{(t+1)} = \frac{1}{n\pi_j^{(t+1)}} \sum_{i=1}^{n} p^{(t+1)}(j|\boldsymbol{x}_i)\boldsymbol{x}_i,$$
$$\sum_{j}^{(t+1)} = \frac{1}{n\pi_j^{(t+1)}} \sum_{i=1}^{n} p^{(t+1)}(j|\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_j^{(t+1)})^T \tag{6}$$

収束条件

設定した十分小さな値 $c > 0$ に対し，観測データに基づく尤度関数について，式 (7) が c 以下なるまで反復更新する．

$$\left| \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{g} \pi_j^{(t+1)} f_j(\boldsymbol{x}_i|\boldsymbol{\mu}_j^{(t+1)}, \sum_{j}^{(t+1)} \right\} - \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{g} \pi_j^{(t)} f_j(\boldsymbol{x}_i|\boldsymbol{\mu}_j^{(t)}, \sum_{j}^{(t)} \right\} \right| \tag{7}$$

さまざまなモデルの候補の中で，どのモデルが最も良くデータ発生の確率構造を捉えているかを決定する必要がある．このような問題に対しては，$g = 2, 3, \cdots$ と設定して EM 法によって推定した混合分布モデルを情報量基準 AIC, あるいはベイズ評価基準 BIC によって評価，選択する方法が考えられる．本研究では BIC を用いてモデルの評価を行った．

$$BIC = -2 \ln L + k \ln n \tag{8}$$

L は尤度関数，k はモデルのパラメータ個数，n は標本の大きさを表している．

3.3 ラインアップ分析

ソフトクラスタリングによって割り当てられた各クラスタを新ポジションとし，各クラスタに所属する確率を用いたラインナップ分析をおこなう．表 5 のようにクラスタ確率を基にラインアップを作成し，各ポジションのクラスタ確率の合計値を説明変数とする．目的変数は Off RTG と Def RTG を使用する．作成したラインナップデータに対し，LightGBM を用いた決定木による回帰分析を行う．

実際にラインアップを構築する際に，選手のポジションだけでなく，各選手のスキルも考慮する必要がある．そこで，選手のスキルとクラスタ確率の両方を使用し，クラスタ確率のみ使用した場合と予測精度を比較する．また，各チームは 2, 3 人のスター選手を中心にラインアップを構築しており，実際に予測モデルを用いてスター選手を中心とした選手の組み合わせを最適化する．

表 5: クラスタ確率を用いたラインナップの例

選手名	Cluster 1	Cluster 2	Cluster 3	$\cdots$	Cluster n
選手 1	0.30	0.70	0.00	$\cdots$	0.00
選手 2	0.00	1.00	0.00	$\cdots$	0.00
選手 3	0.10	0.00	1.00	$\cdots$	0.00
選手 4	1.00	0.00	0.00	$\cdots$	0.00
選手 5	0.00	0.00	0.00	$\cdots$	1.00
合計値	1.40	1.70	1.00	$\cdots$	1.00

3.3.1 LightGBM

勾配ブースティング決定木 (以下： GBDT) は有名な機械学習アルゴリズムであり，XGBoost や pGBRT 等多くの効果的な実装が存在する．それらの実装の多くは特徴量の次元が高く，データサイズが大きい場合，各特徴量について可能性のある全ての分割点の情報利得の推定のために，全てのデータインスタンスをスキャンする必要があり，非常に時間がかかる．LightGBM では，勾配片側サンプリング (以下：GOSS) と専用機能バンドル (以下：EFB) を用いる事で，従来の GBDT と同等の精度を達成しながら，学習プロセスを最大 20 倍高速化することが示された．

3.3.2 Gradient Boosting Decision Tree

GBDT は決定木のアンサンブルモデルであり，決定木の学習において最適な分割点を見つけることが最も時間を要する．最適な分割点を見つけるためのアルゴリズムとして，1 に示すようにヒストグラムに基づくアルゴリズムがある．ヒストグラムベースアルゴリズムはソートされた特徴量から分割点を見つけるのではなく，連続的な特徴量をビンにし，学習時にこれらのビンを使用して特徴量のヒストグラムを構築する．ヒストグラムベースアルゴリズムは，メモリ消費量と学習速度の両方で効率的である．

Input: I: training data, d: max depth
Input: m: feature dimension
$nodeSet \leftarrow \{0\}$ ▷ tree nodes in current level
$rowSet \leftarrow \{\{0, 1, 2, ...\}\}$ ▷ data indices in tree nodes
for $i = 1$ ***to*** d **do**
  for $node$ ***in*** $nodeSet$ **do**
    usedRows $\leftarrow rowSet[node]$
    for $k = 1$ ***to*** m **do**
      $H \leftarrow$ new Histogram()
      ▷ Build histogram
      for j ***in*** *usedRows* **do**
        bin $\leftarrow I$.f[k][j].bin
        H[bin].y $\leftarrow H$[bin].y + I.y[j]
        H[bin].n $\leftarrow H$[bin].n + 1
      Find the best split on histogram H.
      ...
  Update $rowSet$ and $nodeSet$ according to the best split points.
  ...

図 1: Histogram-based Algorithm

Input: I: training data, d: iterations
Input: a: sampling ratio of large gradient data
Input: b: sampling ratio of small gradient data
Input: $loss$: loss function, L: weak learner
models $\leftarrow$ {}, fact $\leftarrow \frac{1-a}{b}$
topN $\leftarrow$ a $\times$ len(I) , randN $\leftarrow$ b $\times$ len(I)
for $i = 1$ ***to*** d **do**
  preds $\leftarrow$ models.predict(I)
  g $\leftarrow loss(I,$ preds), w $\leftarrow$ {1,1,...}
  sorted $\leftarrow$ GetSortedIndices(abs(g))
  topSet $\leftarrow$ sorted[1:topN]
  randSet $\leftarrow$ RandomPick(sorted[topN:len(I)], randN)
  usedSet $\leftarrow$ topSet + randSet
  w[randSet] $\times =$ fact ▷ Assign weight $fact$ to the small gradient data.
  newModel $\leftarrow$ L(I[usedSet], $-$ g[usedSet], w[usedSet])
  models.append(newModel)

図 2: Gradient-based One-Side Sampling

図 1 に示すように，ヒストグラムベースアルゴリズムは特徴量のヒストグラム基づいて最適な分割点を導出する．ヒストグラムの構築には $O(\#data \times \#feature)$, 分割点の導出には $O(\#bin \times \#feature)$ のコストがかかる．通常 $\#bin$ は $\#data$ よりはるかに小さいため，ヒストグラムの構築が計算量の大部分を占める，$\#data$ や，$\#feature$ を減らすことで GBDT の学習を大幅に高速化することができる．

3.3.3 Gradient-based One-Side Sampling

GBDT における各データインスタンスの勾配は，データサンプリングに有用な情報を提供することができる．つまり，あるインスタンスで勾配が小さい場合，学習誤差は小さく十分に学習できていることがわかる．これらの十分に学習されたデータインスタンスを削除することが考えられるが，データ分布が変化してしまい，学習済みモデルの精度が損なわれてしまう．この問題を回避するために GOSS という新たな手法を用いる．

GOSS は勾配が大きいインスタンスを全て残し，勾配が小さいインスタンスに対してランダムサンプリングを行う．情報利得を計算する際，データ分布への影響を補正するために，GOSS は勾配の小さいデータインスタンスに対して定数 $\frac{1-a}{b}$ で増幅させる (図 2 参照)．具体的には，GOSS はまずデータインスタンスを勾配の絶対値に従ってソートし，上位 $a \times 100\%$ のインスタンスを選択する．次に，残りのデータから $b \times 100\%$ のインスタンスをランダムサンプリングする．その後，GOSS は情報量を計算する際に，サンプリング

された勾配の小さなデータを定数 $\frac{1-a}{b}$ で増幅する．それにより，元のデータ分布をあまり変えずに学習不足のインスタンスに重点を置く．

GBDT は決定木を用いて入力空間 X^s から勾配空間 G への関数を学習する．ここで，n 個の独立同一分布インスタンス $\{x_1, \cdots, x_n\}$ からなる学習データがあり，各 x_i は空間 X^s における s 次元のベクトルであるとする．勾配ブースティングの各反復において，モデルの出力に関する損失関数の負の勾配は $\{g_1, \cdots, g_n\}$ と表記される．決定木モデルは最も情報量の多い特徴で各ノードを分岐させる．GBDT では，情報利得は通常分割後の分類で測定され，式 (9) で定義される．なお，決定木の固定ノードに関する学習データを O とする．式 (9) はこのノードの点 d で分割したときの分散利得を表す．

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{(\sum_{\{x_i \in O : x_{ij} \leq d\}} g_i)^2}{n^j_{l|O}(d)} + \frac{(\sum_{\{x_i \in O : x_{ij} > d\}} g_i)^2}{n^j_{r|O}(d)}\right) \tag{9}$$

ここで，$n_O = \sum I[x_i \in O], n^j_{l|O}(d) = \sum I[x_i \in O : x_{ij} \leq d]$ かつ $n^j_{r|O}(d) = \sum I[x_i \in O : x_{ij} > d]$ である．特徴量 j に対し，決定木アルゴリズムは $d^*_j = argmax_d V_j(d)$ を選択し，最大の利得 $V_j(d^*_d)$ を計算する．次に，点 d_{j*} にける特徴 j^* に従い，データを左右の子ノードに分割する．

提案する GOSS 法では, まず学習インスタンスをその勾配の絶対値に従って降順にランク付けする. 次に，より大きな勾配を持つ上位 $a \times 100\%$ のインスタンスを残し, インスタンス部分集合 A を得る．より小さな勾配を持つ $(1-a) \times 100\%$ のインスタンスからなる残りの集合 A^c に対して，さらにサイズ $b \times |A^c|$ の部分集合 B をランダムサンプリングする．最後に，部分集合 $A \cup B$ に対する推定分散利得 $\tilde{V}_j(d)$ に従いインスタンスを分割し，式 (10) のようになる．

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i)^2}{n^j_l(d)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i)^2}{n^j_r(d)}\right) \tag{10}$$

ここで，$A_l = \{x_i \in A : x_{ij} \leq d\}, A_r = \{x_i \in A : x_{ij} > d\}, B_l = \{x_i \in B : x_{ij} \leq d\}, B_r = \{x_i \in B : x_{ij} > d\}$ であり，係数 $\frac{1-a}{b}$ は B の勾配の総和を A^c のサイズに戻して正規化するために用いる．

GOSS は全インスタンスに対する正確な $V_j(d)$ ではなく，より小さなインスタンス部分集合に対する推定 $\tilde{V}_j(d)$ を用いて分割点を決定するため，計算コストを大幅に削減することが可能である．さらに重要なことは，定理 1 により，GOSS は学習精度をあまり落とさずランダムサンプリングより優れていることが示されることである．

定理 1. GOSS における近似誤差を $\varepsilon(d) = |\tilde{V}_j(d) - V_j(d)|$ かつ $\bar{g}^j_l(d) = \frac{\sum_{x_i \in (A \cup A^c)_l} |g_i|}{n^j_l(d)}, \bar{g}^j_r(d) = \frac{\sum_{x_i \in (A \cup A^c)_r} |g_i|}{n^j_r(d)}$ とする．少なくとも $1-\delta$ の確率で式 (11) が成り立つ.

$$\varepsilon(d) \leq C^2_{a,b} \ln \frac{1}{\delta} \cdot max\left\{\frac{1}{n^j_l(d)}, \frac{1}{n^j_r(d)}\right\} + 2DC_{a,b}\sqrt{\frac{\ln \frac{1}{\delta}}{n}} \tag{11}$$

ただし, $C_{a,b} = \frac{1-a}{\sqrt{b}} max_{x_i \in A^c} |g_i|$, かつ $D = max(\bar{g}^j_l(d), \bar{g}^j_r(d))$

3.3.4 Exclusive Feature Bundling

高次元データは通常非常に疎らである．疎らな特徴量空間では，多くの特徴量が互いに排他的であり，同時にゼロ以外の値をとらない．このような排他的な特徴量を 1 つの特徴量にまとめることを Exclusive Feature Bundling (以下：EFB) と呼ぶ．特徴量走査アルゴリズムにより，個々の特徴量から得られるものと同じヒストグラムを特徴量バンドルから構築できる．このように，$\#bundle << \#feature$ とすることで，ヒストグラム構築の複雑さは $O(\#data \times \#feature)$ から $O(\#data \times \#bundle)$ へと変化する．これにより，精度を落とすことなく GBDT の学習を大幅に高速化することができる．

EFB アルゴリズムでは，多くの排他的特徴量をより少ない密な特徴量にバンドルできるため，特徴量がゼロの場合の不要な計算を効果的に回避することができる．各特徴量にゼロ以外のデータを記録するテーブルを使用することで，基本的なヒストグラムベースアルゴリズムを，ゼロの特徴量を無視する方向に最適化することも可能である．このテーブルをスキャンすることで，1 つの特徴量に対するヒストグラム構築のコストは $O(\#data)$ から $O(\#non\_zero\_data)$ へと変化する．しかし，この方法では木構造全体の成長過程において，これらの特徴量ごとのテーブルを維持する為に，追加のメモリと計算コストが必要となる．そこで，LightGBM ではこの最適化を基本機能として実装した．

4 結果

4.1 新ポジションの提案

4.1.1 クラスタリング結果

統計解析ソフトの R により，mclust パッケージを使用してソフトクラスタリングを行った．ハードクラスタリングではクラスタ数を事前に決めるのに対し，ソフトクラスタリングはクラスタ数とそれに所属する確率を求めることができる．クラスタリングの結果，オフェンスは 7 つ，ディフェンスは 5 つのクラスタとなった．それぞれの分布を図 3，4 に示す．

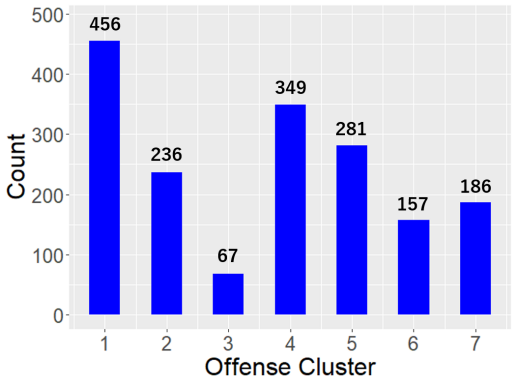

図 3: オフェンスクラスタの分布

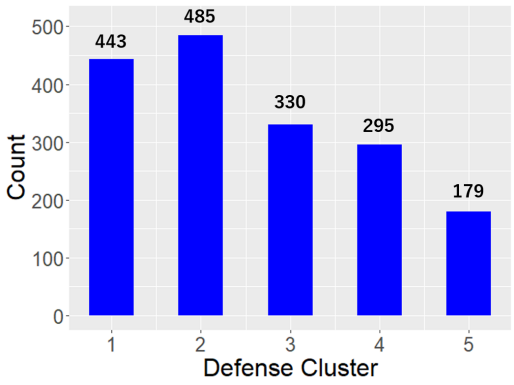

図 4: ディフェンスクラスタの分布

各クラスタのスタッツを比較し，特徴を表 6, 8 にまとめ新ポジションを定義した．また，表 7，9 にはポジションに該当する選手を例として示す．

表 6: 新ポジションの定義 (オフェンス)

新ポジション	特徴	ハイスコア スタッツ	ロースコア スタッツ
Solid Guard	ドリブルからのプレー展開が多く，アイソレーションが少ないアシスト重視のガード.	PUL DRI, AST	CAT 2FGS% OF REB, RES
Utility Forward	アウトサイドシュートの成功率が高く，インサイドでも得点できる汎用性の高いフォワード.	MID FGA% USG, 3P%	RES, AST OF REB
High Usage Forward	ポストアップが多く，USG が高い. 3P% が低く，ミドルレンジやインサイドでオフェンスを展開する.	MID FGA% POS, USG	3P FGA%，FT% CAT 3FGA%
Spot Up Shooter	3 ポイントシュートを得意とし，キャッチアンドシュートが多く，タッチ数やドライブ数は少ない.	CAT 3FGA% C3 FGA% 3P%, SPO	FGA，USG ITP FGA%
Stretch Center	インサイドで得点やリバウンドをこなし，3 ポイントシュートの確率は平均以上の値を持つ.	OF REB 3P%，POS RES FGA%	PUL FGA%，TRN AST
Traditional Center	リバウンドとインサイドの得点割合が最も高く，3 ポイントシュートはほとんど打たない.	OF REB，FTA% RES FGA%	3P%，TOUCH TRN
Scoring Guard	Solid Guard よりも USG が高く，シュート本数や，アイソレーションが多い.	FGA，USG AST, ISO	OF REB C3 FGA% CAT 3PFGA%

表 7: 新ポジションの該当選手紹介 (オフェンス)

新ポジション	例
Solid Guard	’16 Patty Mills ’20 Fred VanVleet
Utility Forward	’16 Klay Thompson ’19 Kawhi Leonard
High Usage Forward	’19 Giannis Antetokounmpo ’20 Joel Embiid
Spot Up Shooter	’17 PJ Tucker ’20 Danny Green
Stretch Center	’16 Kevin Love ’18 Draymond Green
Traditional Center	’18 Jusuf Nurkic ’20 Rudy Gobert
Scoring Guard	’16 Stephen Curry ’20 Mike Conley

表 6 から, 同じセンターポジションにおいて, 従来のリング周辺でプレーする Traditional Center とアウトサイドのシュートもこなす Stretch Center の 2 種類のポジションを作成することができた．通常ディフェンスにおいては，背の高い選手がより多くラインアップにいた方が有効だが，センターばかりを揃えるとオフェンスでスペースをお互いにつぶし

あってしまう．しかし，ラインナップにセンターが 2 人いたとしても，Traditional Center と Stretch Center の組み合わせの場合，互いに異なる特徴を持つため，コートのスペースを狭くすることなくプレーできる可能性が示された．3 ポイントシュートが得意な選手をシューターと呼ぶが，ガードの役割もこなす Scoring Guard と，シュートを主軸とする Spot Up Shooter に分けることで，ガードとシューターの両方の特徴を持つポジションが提案された．試合中の細かなプレーに関するスタッツを使用することで，従来よりもポジション細分化することができた．

表 8: 新ポジションの定義 (ディフェンス)

新ポジション	特徴	ハイスコア スタッツ	ロースコア スタッツ
Average Outside Defender	平均的なスタッツを持ち，移動 スピードは平均以上の値を持つ．	SPE	CON 2P BLK
Aggressive Defender	平均的な身長で，3 ポイント シュートに対し素早く シュートチェックにいく回数が多い．	SPE，CON 3P	BLK，DF REB
Rim Protector	ブロック数が最も多く， リング周辺でディフェンス するため移動速度は遅い．	DF REB，BLK CON 2P	STL，DFL CON 3P
Versatile Defender	スティール以外のスタッツは 平均的に高く，汎用性が高い．	dfISO，DF REB CON 3P	STL，DFL
Elite Outside Defender	スティールやディフレクションが 多く，アウトサイドディフェンス を得意とする．	STL，DFL SPE	BLK，dfPOS CON 2P

表 9: 新ポジションの該当選手紹介 (ディフェンス)

新ポジション	例
Average Outside Defender	’18 Stephen Curry ’20 Mike Conley
Aggressive Defender	’16 Klay Thompson ’20 Paul George
Rim Protector	’19 Brook Lopez ’20 Rudy Gobert
Versatile Defender	’18 Draymond Green ’18 Pascal Siakam
Elite Outside Defender	’17 Andre Roberson ’20 Kawhi Leonard

ディフェンスでは，ブロックが得意な高身長の Rim Protector や，スティールやディフレクションが得意な低身長の Elite Outside Defender，汎用性の高い Versatile Defender 等，異なる特徴を持つポジションが提案された．ディフェンスにおいて重視される BLK や STL 以外にも，ディフェンス時のスピードも考慮したことで，素早い動きとシュートチェックを得意とする Aggressive Defender を作成することができた．

各オフェンスとディフェンスの特徴量の一部を図 5，6 に示す．標準化されたスタッツの中で，0 以上のスタッツはオレンジに，0 未満のスタッツは灰色に塗りつぶした．Stretch

Center はリバウンドやポストアップと共に，3P% も平均以上の値を持つ．Versatile Defender はスティール以外のスタッツが平均的に高い値を持つ．この 2 つのポジションに該当する選手として，2020-21 シーズンの Draymond Green が挙げられるが，オフェンスとディフェンスの両方で汎用性の高い選手であることが示された．実際の所属チームでも，様々な役割をこなしていることから，選手の特徴をオフェンスとディフェンスの両方から特定できることが示された．

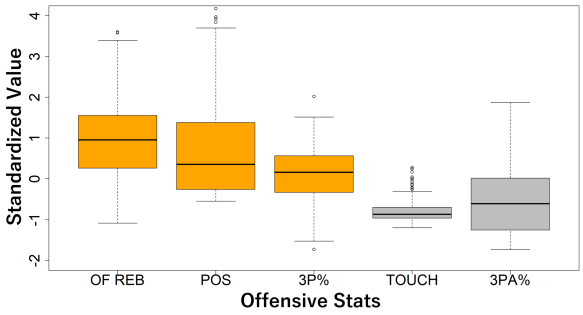

図 5: Stretch Center のスタッツプロット図

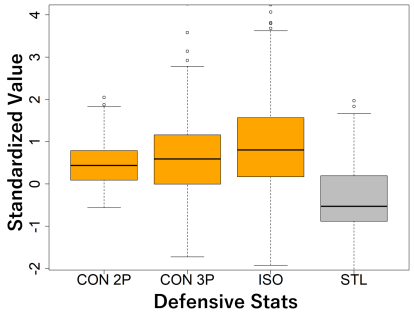

図 6: Versatile Defender のスタッツプロット図

4.1.2　複数ポジション所属選手

多くの選手は 1 つのクラスタに所属する確率が高いが，2 つ以上のクラスタに所属する選手も存在する．ソフトクラスタリングを用いる事により，複数ポジションの役割をこなす選手を定量的に示すことができた．実際に複数クラスタに所属する選手として以下のような選手が挙げられる．オフェンスでは，2018-19 シーズンの Mike Conley は Solid Guard 45%, Scoring Guard 55% の割合となり，バランス良くガードの役割をこなせる万能選手であることがわかる．ディフェンスでは，2017-18 シーズンの Giannis Antetokounmpo は Rim Protector 36%, Versatile Defender 64% の割合となり，インサイドとアウトサイド両方を守りながら，ブロックやディフェンスリバウンドをこなせるオールラウンドなディフェンダーであることがわかる．

クラスタリングによりオフェンスは 7, ディフェンスは 5 つのポジションが提案されため，これらを組み合わせることで，最大 35 通りの組み合わせ，つまり 35 通りの選手の特徴が得られた．通常のオフェンスとディフェンスを合わせた 5 通りのポジションによる選手の分類に比べ，より細かく分類することができた．なお，実際の選手データから得られた組み合わせは 30 通りとなった．

4.2　ラインナップ分析

4.2.1　ラインナップ予測モデルの構築と検証

表 10 に示すようにクラスタ確率を基にラインアップを組み，OffRTG と DefRTG を予測するモデルを構築した．表 10 のデータは 2018-19 シーズン Sacramento Kings のラインナップデータである．

表 10: 2018-19 シーズン Sacramento Kings 所属選手のクラスタ確率 (オフェンス)

選手名	Solid Guard	Utility Forward	High Usage Forward	Spot Up Shooter	Stretch Rebounder	Traditional Center	Scoring Guard
I. Shumpert	0.73	0.00	0.00	0.27	0.00	0.00	0.00
W. Cauley-Stein	0.00	0.00	0.00	0.00	0.00	1.00	0.00
B. Hield	0.02	0.00	0.00	0.00	0.00	0.00	0.98
D. Fox	1.00	0.00	0.00	0.00	0.00	0.00	0.00
M. Bagley III	0.00	0.00	0.00	0.00	1.00	0.00	0.00

ラインナップデータは各ラインアップにおける出場時間の差が非常に大きく，ノイズが多い傾向にある．例として，あるラインアップがシーズン通算 3 分出場で，その間のスコアが 10 - 0 でリードしていた場合，実際の OffRTG，DefRTG よりも良く評価され，正当な評価ができないという問題がある．そこで，先行研究と同様に，600 ポゼッションを基準とした調整済み OffRTG，DefRTG を使用する．600 ポゼッションとは約 6 試合分のポゼッションであり，6 試合未満の場合は各スタッツや得失点において極端な値をとることがあり，それを除外するためである．式 (12) に調整済み Rating の計算方法を示す．式 (12) は，600 ポゼッション未満のラインアップは OffRTG，DefRTG が調整されていることを示している．Team OffRTG，Team DefRTG はチームのシーズン平均値を示している．

$$
\begin{aligned}
&\frac{Possessions}{600} \geq 1, \quad R_{Off} = OffRTG,\ R_{Def} = DefRTG \qquad (12)\\
&\frac{Possessions}{600} < 1, \quad R_{Off} = \frac{Possessions}{600} \times OffRTG + \left(1 - \frac{Possessions}{600}\right) \times Team\ OffRTG\\
&\qquad\qquad R_{Def} = \frac{Possessions}{600} \times DefRTG + \left(1 - \frac{Possessions}{600}\right) \times Team\ DefRTG
\end{aligned}
$$

5 シーズン分のラインナップデータから，表 10 に示すクラスタ確率を用いたソフトラインナップを 10,000 件作成した．選手個人のデータ取得条件から，出場時間が 30 試合未満，または 12 分未満の選手を除外した場合，ソフトラインナップデータは 7,634 件となった．表 10 から 2018-19 シーズンの Sacramento Kings のソフトラインナップを例として表 11 に示す．

表 11: 2018-19 シーズン Sacramento Kings ソフトラインナップ (オフェンス)

Solid Guard	Spot Up Shooter	Stretch Rebounder	Traditional Center	Scoring Guard
1.75	0.27	1.00	1.00	0.98

表 11 に示されていないポジションは確率 0 である．これら 7 つのクラスタ確率を説明変数とし，オフェンスとディフェンスそれぞれ OffRTG, DefRTG を目的変数として LightGBM による回帰分析を行った．新ポジションと目的変数の関係性を分析するにあたり，それらの関係性を非線形に表し，ポジション間の相互作用を考慮する為に LightGBM を使用した．予測精度の評価方法として，分析用データセットを学習データと検証用データに 4 : 1 の割合で分割し，平均平方二乗誤差 (RMSE) により評価する．式 (13) 中の n はデータ数，$\hat{y}$ は予測値，y は正解値とする．予測モデルの構築のために，確率の間隔を

0.5 とし，全組み合わせのラインナップデータを作成したものを表 12 に示す．

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{13}$$

表 12: 予測用データセット (オフェンスクラスタ確率)

Solid Guard	Utility Forward	High Usage Forward	Spot Up Shooter	Stretch Rebounder	Traditional Center	Scoring Guard
5.00	0.00	0.00	0.00	0.00	0.00	0.00
4.50	0.50	0.00	0.00	0.00	0.00	0.00
4.00	0.50	0.50	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...
0.00	0.00	0.00	0.00	0.00	0.00	5.00

回帰分析の精度は，RMSE がオフェンスでは 2.09，ディフェンスでは 2.14 となり，どちらも低く，誤差は小さい値となった．分析結果から，予測値が最も高いラインナップを表 13，14 に示す．表 13 から，OffRTG を高めるには Scoring Guard 2 人分，Stretch Center 1.25 人分と，Solid Guard，Spot Up Shooter，Traditional Center が約 0.5 人分づつとなる組み合わせが最適なラインナップとなった．現在 NBA ではガードが攻撃のリーダーとなることが多く，予測モデルにおいても Scoring Guard がオフェンスにおいて重要であることがわかる．

表 13: 調整済み Offensive Rating 予測値上位 5 位のラインナップ (クラスタ確率使用)

Solid Guard	Utility Forward	High Usage Forward	Spot Up Shooter	Stretch Center	Traditional Center	Scoring Guard	予測値 (OFF RTG)
0.50	0.50	0.00	0.50	1.25	0.25	2.00	117.67
0.50	0.25	0.00	0.75	1.25	0.25	2.00	117.67
0.50	0.50	0.00	0.00	1.25	0.75	2.00	117.46
0.50	0.50	0.00	0.75	1.25	0.00	2.00	117.46
0.50	0.50	0.00	0.75	0.50	0.75	2.00	117.45

表 14 から，DefRTG を高めるには，Rim Protector，Elite Outside Defender が 2 人分ずつと，Average Defender 1 人分の組み合わせが最適なラインナップとなった．ディフェンスではブロックやスティールを得意とするポジションを多く起用することで DefRTG を高めることが示された．Rim protector と Elite Outside Defender が 2 人分ずつ配置できるのであれば，Aggressive Defender や Versatile Defender をラインナップに配置する重要性は高くない．しかし，Elite Outside Defender に所属される選手の数は 1 チーム平均 1.23 人しかおらず，所属確率が 0.25 以上の選手を含めても 1 チーム平均 1.8 人であり，常にラインナップに 2 人分配置することは困難である．

表 14: 調整済み Defensive Rating 予測値上位 5 位のラインナップ (クラスタ確率使用)

Average Outside Defender	Aggressive Defender	Rim Protector	Versatile Defender	Elite Outside Defender	予測値 (DEF RTG)
1.00	0.00	2.00	0.00	2.00	101.45
0.75	0.00	2.00	0.00	2.25	101.68
0.75	0.00	2.25	0.00	2.00	101.96
0.00	1.00	2.25	0.75	1.00	102.02
1.00	0.00	2.25	0.00	1.75	102.02

アウトサイドのシュート確率が高い選手はディフェンスを引き寄せるため，リング周辺のスペースを広げることができる．3 ポイントシュートを得意とする Spot Up Shooter と，アウトサイドシュート確率が良いセンターである Stretch Forward の関係性は，役割が重複しているため図 7 のように予測値は低い値となる．インサイドを攻める High Usage Forward との相性は良く，図 8 の様に Spot Up Shooter に対し，1 人分以上の High Usage Forward であれば高い予測値となった．ディフェンスでは，図 9，10 から， Elite Outside Defender を 1 人分以上とした際に，Aggressive Defender よりも Average Outside Defender を使用した方が良いことがわかる．

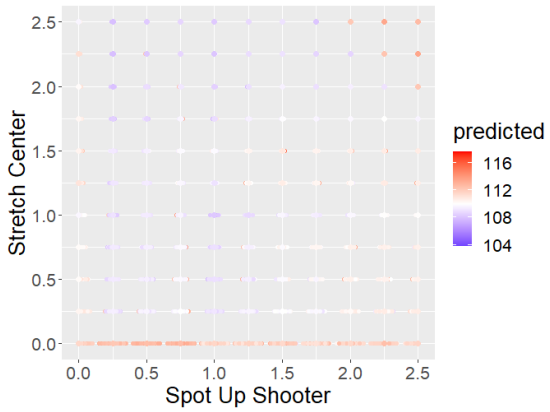

図 7: Spot Up Shooter と Stretch Center の予測値プロット図

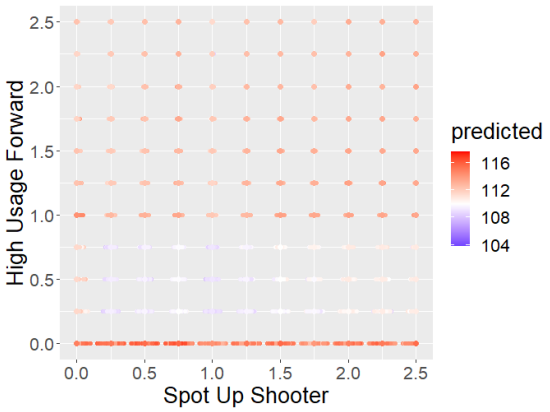

図 8: Spot Up Shooter と High Usage Forward の予測値プロット図

4.2.2 スキルを考慮したラインナップ予測

選手のスキルとクラスタ確率の両方を使用し，クラスタ確率のみ使用した場合と予測精度を比較した．また，実際のチームに対して予測モデルを用いてラインアップを構築する．

選手のスキルの評価指標として，チームの評価指標として使用した OffRTG，DefRTG の選手版である Player OffRTG，Player DefRTG を使用する．Player OffRTG, DefRTG は該当選手が出場している際のチームの 100 ポゼッションごとの得点数，失点数となる．前節では説明変数としてクラスタ確率を使用したが，本節では説明変数としてクラスタ確率に Player OffRTG または Player DefRTG を乗じた変数を使用し，目的変数は調整済み OffRTG，DefRTG を使用してソフトラインナップを作成する．また，説明変数の間隔を 30 とし，0 〜 600 までの範囲における全組み合わせのラインナップデータを作成したものを表 15 に示す．

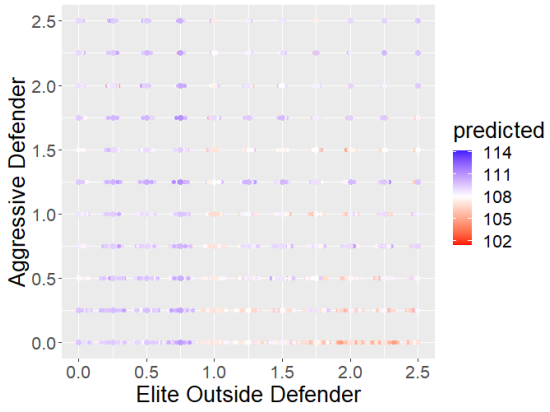

図 9: Elite Outside Defender と Aggressive Defender の予測値プロット図

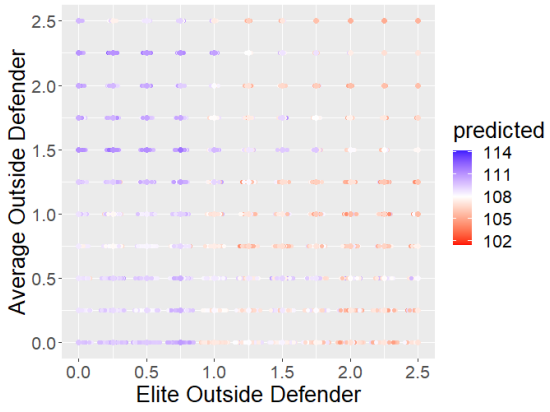

図 10: Elite Outside Defender と Average Outside Defender の予測値プロット図

表 15: 予測用データセット (クラスタ確率 ×Player OFF RTG)

Solid Guard	Utility Forward	High Usage Forward	Spot Up Shooter	Stretch Rebounder	Traditional Center	Scoring Guard
600	0.00	0.00	0.00	0.00	0.00	0.00
570	30	0.00	0.00	0.00	0.00	0.00
540	30	30	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...
0.00	0.00	0.00	0.00	0.00	0.00	600

回帰分析の精度は，RMSE がオフェンスでは 1.96，ディフェンスでは 1.86 となり，どちらも低く誤差は小さい値となった．表 15 に示した全組み合わせデータを予測モデルに当てはめ，予測値を得た．分析結果から，予測値が最も高いラインナップを表 16，17 に示す．表 16 から，調整済み OFF RTG を高めるには Solid Guard, High Usage Forward, Traditional Center の値がそれぞれ 120，Scoring Guard の値が 150 となるようにラインアップを構築することが最適であることが示された．クラスタ確率のみを使用した場合と比べ，最適なラインナップにおいて Spot Up Shooter が採用され，Stretch Center の値は 0 となった．

表 16: 調整済み Offensive Rating 予測値上位 5 位のラインナップ (クラスタ確率 ×Player OFF RTG 使用)

Solid Guard	Utility Forward	High Usage Forward	Spot Up Shooter	Stretch Center	Traditional Center	Scoring Guard	予測値 (OFF RTG)
150	30	0.00	120	0	120	120	117.18
150	30	0.00	120	0	150	120	117.18
150	60	0.00	120	0	120	120	117.16
150	60	0.00	120	0	150	120	117.16
30	30	0.00	270	150	0	120	117.13

ディフェンスでは，表 17 のように Rim Protector の値以外 0 という極端な結果となってしまった．実際の試合において同じポジションの選手のみでラインアップを構築することは考えにくく，スキルを考慮した場合のポジション間の相互作用を十分に示すことができなった．

表 17: 調整済み Defensive Rating 予測値上位 5 位のラインナップ (クラスタ確率 ×Player DEF RTG 使用)

Average Outside Defender	Aggressive Defender	Rim Protector	Versatile Defender	Elite Outside Defender	予測値 (DEF RTG)
0	0	450	0	0	75.32
0	0	480	0	0	75.32
0	0	510	0	0	75.32
0	0	540	0	0	75.32
0	0	570	0	0	75.32

次に，実際のチームを対象に予測モデルを用いてオフェンスラインアップを構築する．対象とするチームは 2020-21 シーズンの Los Angeles Lakers とし，チームを代表するスター選手である Lebron James と Anthony Davis を中心選手として固定する．ラインアップ構築の条件として，同シーズンの全チームの選手を対象とし，そのうち同シーズンのオールスターに選出された選手は対象外とした．仮に，スター選手ばかりをラインナップに採用した場合，予測値が高くなるのは当然であり，実際のチームにはスター選手は約 2 人程度しか在籍していないためである．上記 2 選手の OFF RTG を使用し，予測モデルにおいて最も高い値となるように選手を選択すると，表 18 に示すようなラインアップとなった．実際のシーズン中に Los Angeles Lakers において最も高い OffRTG となったラインアップデータの値は 109.80 であったため，予測モデルを使用して構築されたラインアップの方が，100 ポゼッションあたり 3.38 点多く得点することが示された．

表 18: 2020-21 シーズン Los Angeles Lakers のラインアップ構築

選手 1	選手 2	選手 3	選手 4	選手 5	予測値 (OFF RTG)
L. James	A. Davis	M. Brogdon	J. Brown	J. Collins	113.18

5 考察

5.1 まとめ

本論文では，ポジションをオフェンスとディフェンスに分けてクラスタリングを行う事で，それぞれ特徴の異なる新ポジションの提案を目指し，クラスタリング分析を行った．クラスタリングの結果，オフェンスとディフェンスどちらにおいても従来のポジションに比べ，より選手の特徴を詳細に示すことができた．Stretch Center のような特徴を持つ選手は汎用性が高く，現代の NBA では重要なポジションであるが，Traditional Center や Solid Guard のような従来のポジションの特徴に近い場合でも，OffRTG に効いていることが示された．

選手のスキルをクラスタ確率と共に使用した場合，クラスタ確率のみ使用した場合と異なる結果となった．これは，クラスタ確率のみではポジションのおよその相性しか示せておらず，スキルの上下を考慮することで相性が変化したためであると考えられる．また，ディフェンスではスキルを考慮した場合，ラインナップに Rim Protector を 4, 5 人分取り入れることで DefRTG が最も良い予測値となってしまった．これは，ディフェンスポジ

ション間の相互作用を正確に示すことができなかったためであると考えられる．オフェンスラインナップの作成では，予測モデルに基づき，実際のチームから新たに最適なラインナップを作成し，OffRTG を上昇させることができた．これにより，選手の移籍やチーム戦術の提案にも繋がる結果が得られた．

5.2 課題

最後に本論文の課題を述べる．ディフェンスにおいて，スキルとクラスタ確率を使用した予測モデルの構築では，扱うスタッツの数が 11 個と少なかったため十分にディフェンススキルを評価できなった．それにより，ポジション間の相互作用が示されず，極端な値の予測モデルとなってしまった．より詳細なディフェンスのスタッツを追加することで精度の向上につながると考えられる．

本論文では，オフェンスとディフェンスに分けてポジションを提案したが，それらを組み合わせて 1 つの予測モデルを作成することができれば，より実戦的で使いやすくすることができると考えられる．

謝辞

本研究を進めるにあたり，ご指導いただきました横浜市立大学 Micheletto Ruggero 先生には心よりお礼申し上げます．

参考文献

[1] BasketBall Positions (2015) NBA.com.
< https://jr.nba.com/basketball-positions/#: :text=Players%20in%20a%20basketball%20game,point%20guard%2C%20and%20shooting%20guard.&text=The%20center%20is%20the%20tallest,on%20close%20shots%20and%20rebound. > 2021. 5. 10.

[2] Samuel Kalman, Jonathan Bosh (2020), *NBA Lineup Analysis in Clustered Player Tendencies: A new approach to the positions of basketball & modeling lineup effiency of soft lineup aggregates*, MIT SLOAN SPORTS ANALYTICS CONFERENCE.
< https://www.sloansportsconference.com/research-papers/nba-lineup-analysis-on-clustered-player-tendencies-a-new-approach-to-the-positions-of-basketball-modeling-lineup-efficiency. > 2021. 2. 9.

[3] Player General, NBA.com.
< https://www.nba.com/stats/players/traditional/. > 2020. 10. 22.

[4] Player Playtype, NBA.com.
< https://www.nba.com/stats/players/isolation/. > 2020. 10. 22.

[5] Player Tracking, NBA.com.
< https://www.nba.com/stats/players/drives/. > 2020. 10. 22.

[6] Player Opponent Shooting, NBA.com.
< https://www.nba.com/stats/players/opponent-shooting/> 2020. 10. 22.

[7] Player Hustle, NBA.com.
< https://www.nba.com/stats/players/hustle/> 2020. 10. 22.

[8] Lineups, NBA.com.
< https://www.nba.com/stats/lineups/traditional/. > 2020. 10. 25.

[9] Team General, NBA.com.
< https://www.nba.com/stats/teams/advanced/. > 2020. 10. 22.

[10] 松井秀俊 · 小泉和之 (2019) 統計モデルと推測，講談社.

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu (2017), *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, Neural Information Processing Systems
< https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. > 2020. 10. 9.