LightGBM を用いた各プレー場面におけるシュート予測とその要因分析

佐藤洸太朗・東京理科大学理工学部経営工学科
澤田智秀・東京理科大学理工学部経営工学科
東瀬皓太郎・東京理科大学理工学部経営工学科
田中耀大・東京理科大学理工学部経営工学科
藪谷瑞生・東京理科大学理工学部経営工学科
安井清一・東京理科大学理工学部経営工学科
〒278-8510 千葉県野田市山崎 2641
TEL: 04-7122-9288, E-mail: yasui@rs.tus.ac.jp

1. 背景目的

サッカーの試合観戦はシュートなどのゴール付近のプレー以外では，その試合が白熱しているかどうか，つまり現時点のプレーがいかにしてシュート･得点につながるかが不透明であり，場面状況を正確に解釈することは難しい．特にサッカー経験者以外では顕著である．これは野球などのターン制スポーツとは異なり，場面が連続的かつ激しく変化することに起因すると考えられる．

本研究では，試合が白熱しているかどうかを視聴者にわかりやすく伝えるために，時々刻々と移り変わる試合状況を機械学習で評価し，10 秒以内にシュートが起こる確率を提示するシステムを提案する．

2. シュートの予測

A. データ説明

本研究では，第 10 回スポーツデータ解析コンペティションのデータスタジアム株式会社から提供されたデータ，2019 年の J1 最終 5 節の 45 試合のデータとプレボールタッチデータを用いた．トラッキングデータは 1/25 秒ごとの各選手とボールの位置座標データであり，ボールタッチデータは 1/30 秒ごとの各アクション(121 項目)についてのデータである．

B. 説明変数の作成と目的変数

説明変数はトラッキングデータをもとに試合場面の状況を示す様々な特徴量として定義する．目的変数はボールタッチデータをもとに，ある時点から 10 秒以内にシュートが生じていれば “1”，生じていなければ “0” の 2 値データとする．トラッキングデータとボールタッチデータの 1 秒間あたりのフレーム数が異なっているため，説明変数は 1 秒ごとに平均をとってまとめた．

説明変数となる特徴量はトラッキングデータからボールの x, y 座標以外は松岡ら[1]を基にして 78 変数を生成した．具体的には以下のとおりである．

- 半径 1, 3, 5, 10m 以内にいる選手の人数を以下の区分で算出した．
  - 総選手数
  - ホーム・アウェイチームを区別した選手数
  - ホーム・アウェイチームを区別した FW, MF, DF ごとの選手数
- ホーム・アウェイチーム別に算出した FW, MF, DF の各ポジションにおける x 軸と y 軸の長さ．
- ホーム･アウェイチーム別の DF ラインと，敵ゴールラインに最も近い FW 選手の x 座標 (FW ライン)．
- 松岡ら[1]で定義されているホーム・アウェイチームのプレスレベル．プレスレベルの定義は以下の通りである．

ある時間をt，ボールの位置をBp，選手のいる位置をp，ある選手をiとすると，

ある選手のプレスレベルPiはスピードSiと方向Diから

$$Pi = Si(t) \times Di(Bp(t), pi(t), pi(t-0.04))$$

のように定義される．

スピードはトラッキングデータを参照しており，最小値を 0 km/h，最大値を 35.78 km/h として 0〜10 に正規化した値をスピードSiの値とした．方向は，選手の進行方向と選手からボールがある方向との角度θを算出した．ある選手の進行方向は，ある時間tの時の選手の位置 $pi(t)$ と 1 フレーム前(0.04 秒前)の選手の位置 $pi(t-0.04)$ から算出した．また，選手からボールがある方向は，$pi(t-0.04)$とある時間tの時のボールの位置$Bp(t)$ から算出した(図 1)． 最小値を 180°，最大値を 0°として，0〜10 に正規化した値を方向Diとした．

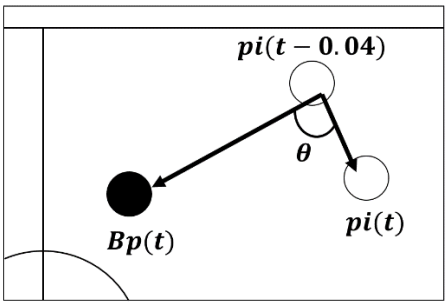

図 1 ：プレスレベル

- コートをペナルティエリアとその左右，そしてそれ以外のセンターラインまでの領域に平行した長さを 3 分割した(図 2)．そして，その 6 領域（センターライン左右で合計 12 領域）でのホームアウェイチームごとの選手数

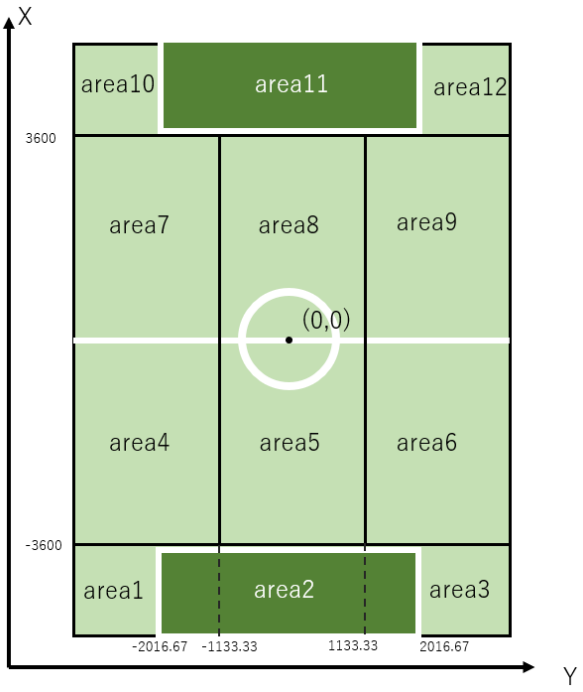

図 2：コートエリア分け定義図

C. データの分割方法

本研究ではトレーニングデータとテストデータの分割を次のように行った．45 試合分のデータをひとまとめにし，約 200,000 行のうち 4 割をランダムに抽出してテストデータに割り当てた．そして残った 6 割のデータをトレーニングデータとした．

D. 予測モデル

2 値分類器として Microsoft 社が提供する勾配ブースティングでの分類器である LightGBM を用いた．学習の評価指標として AUC を使用した．

E. 環境

本研究の機械学習に使用した計算環境は以下の通りである．

- Google Colaboratry
- Scikit – Learn : 0.22.2
- LightGBM : 2.2.3
- Tensorflow：2.30
- Keras：2.4.3

3. 予測結果

予測結果として Accuracy 0.987 (しきい値:0.5) AUC は, 0.988 となった. また, 重要度を LightGBM の feature importance メソッドで算出した (表 1). 重要視されていた要因は「ポジションごとの x, y の長さ」「ポジションごとのライン」であった．逆に重要視されていない要因は「プレスレベル」「ボールの周辺の選手数」「領域の選手数」であった.

4. 考察

本研究を行うに当たって, はじめは使用したデータが時系列データとしてとらえることが可能であると考え，自然言語のニューラルネットワークを用いることを検討した．具体的には自然言語のニューラルネットワークとして，RNN の勾配消失問題を解決し, より長期的な記憶を可能とする LSTM を用いた．しかし，LSTM の AUC が低い水準にとどまり，LightGBM の AUC が高い水準になった．これはサッカーの局面毎の評価は時系列ではなく, 非時系列で捉えられると考えられるためである．今回，目的としたサッカーの局面の客観的な評価を行うには, その時点での選手の位置関係, または位置座標を考慮すればよいといえる. つまり, その局面での情報はそれ以前の局面の動きの結果であり，ある時点の局面のデータというのは時系列のデータを内包しているからであると考えられる.

5. 今後の展望

応用例としては数値化されたシュート確率がホーム・アウェイでそれぞれ予測されるので, 現在の局面においてどちらのチームが優勢かを未経験者でも一目で判断できるシステムの構築につなげられる.

今後の研究としては，今回は決定木の分岐にその特徴量が使われた回数をカウントして求めた重要度(Split Importance)を用いたが，今後はその特徴量で不純度をどれだけ改善できたかという指標で評価した重要度(Gini Importance)を用いた場合との結果の差異を調べる.

また，今回の予測結果と実際の選手の位置座標の変化を合わせた動画を作成し，よりシュートにつながりやすいポジションのパターンを研究する.

6. 参考文献

[1] 松岡弘樹, 田原康寛, 安藤梢, 西嶋尚彦 (2019). “トラッキングデータからの戦術アクション項目の開発”.

7. 謝辞

データを提供してくださった情報・システム研究機構 統計数理研究所 医療機構データ科学研究センター様，データスタジアム株式会社様，皆様にこの場を借りて多大なご協力を心よりお礼申し上げます.

表 1：10 秒以内の予測した際の重要度の上位

説明変数	importance
味方 MF の y の長さ	7687
敵 FW ライン	7545
味方 MF の x の長さ	7225
味方 DF の y の長さ	7160
敵 MF の y の長さ	7104
味方 DF の x の長さ	7044
味方 DF ライン	7030
味方 FW ライン	6926
敵 MF の x の長さ	6902
敵 DF の x の長さ	6304
敵 DF ライン	5815
敵 DF の y の長さ	5815
敵 FW ラインの y の長さ	5319
味方 FW の y の長さ	5081
敵 FW の x の長さ	5026
味方 FW の x の長さ	4936
味方のプレスレベル平均	4376
敵のプレスレベル平均	4104
ボール y 座標	3788
ボール x 座標	3404