Predicting the Receivers of Football Passes

Heng Li[1] and Zhiying Zhang[2]

[1] School of Computing, Queen's University, Kingston, Canada
hengli@cs.queensu.ca
[2] Microsoft, Bellevue, WA, USA
zhiyingz@microsoft.com

Abstract. Football (or association football) is a highly-collaborative team sport. Passing the ball to the right person is essential for winning a football game. Anticipating the receiver of a pass can help football players build better collaborations on the field and help coaches make better tactic decisions. In this work, we use machine learning models to learn how professional football players pass the ball, and leverage the models to predict the receivers of football passes. Our model's first guess, top-3 guesses and top-5 guesses find the correct receiver of a pass with an accuracy of 50%, 84% and 94%, respectively. This paper discusses the features that we use to build our models, our modeling approaches, and the important factors that explain the the receiver of a pass.

Keywords: Football pass prediction · Learning to rank · LightGBM.

1 Introduction

In a football game, players pass the ball to their teammates in order to create good shooting opportunities or prevent the opposing team from getting the control of the ball. Accurately passing ball to the right player is essential for winning a football game [**?**,**?**].

Prior work [**?**,**?**] studies how passing sequences lead to goals. Their findings have shaped the tactics of many football coaches. In this work, we study football players' passing patterns and construct machine learning models to anticipate the receiver of a pass. We believe that football coaches and players can take our results into consideration when they make their tactics or make their passes/runs. Anticipating receivers of passes can also help automatic cameras to always focus on the ball in a game.

This work analyzes a dataset which contains 12,124 passes performed by a Belgian football club in 14 games. We want to answer the following research questions:

RQ1: What are players' passing patterns? We discuss how players pass the ball on the field, e.g., what is their passing accuracy, how often do they pass the ball forwards? Understanding players' passing patterns can help us derive features that can explain the receiver of a pass.

RQ2: How well can we model the receiver of a pass? We construct machine learning models to predict the receiver of a football pass. An accurate model can help coaches and players make informed tactics.

RQ3: What are the influential factors that explain the receiver of a pass? We analyze the models to find the the most influential factors that explain the receiver of a pass. Understanding such influential factors can help coaches and players improve their tactics according to these factors.

Paper organization. The remainder of the paper is organized as follows. Section 2 discusses our approaches for constructing our prediction models, including feature extraction, modeling approaches and evaluation measures. Section 3 present the results for answering our research questions. Finally, Section 4 draws conclusions.

2 Methodology

This section discuss our overall methodology, including our feature extraction process, modeling and evaluation approaches.

2.1 Feature extraction

In this work, we extract four dimensions of features to explain the likelihood of passing the ball to a certain receiver. In total, we extract 58 features. A full list of our features are available at: **[[TODO: link to list of all features]]**. Section 3.3 discusses the most important features for explaining the receiver of a pass.

- **Sender position features.** This dimension of features capture the position of the sender on the field, such as the sender's distance to the other team's goal. We choose this dimension of features because players have different passing strategies at different positions, for example, players may pass the ball more conservatively in their own half but more aggressively in the other team's half.
- **Receiver position features.** This dimension of features capture the position of a candidate receiver, such as the candidate receiver's distance to the sender. Senders always consider candidate receivers' positions when they decide the receiver of a pass.
- **Passing path features.** This dimension of features measure the quality of a passing path (i.e., the path from the sender to a candidate receiver), such as the passing angle. The quality of a passing path can predict the outcome (success/failure) of a pass.
- **Team position features.** This dimension of features capture the overall position of the team in control of the ball, such as the front line of the team. Team position might also impact the passing strategy, for example, a defensive team position might be more likely to pass the ball forwards.

2.2 Removing redundant features

Redundant features usually add more complexity to the model than the information they provide to the model. Redundant features can also result in highly unstable models [**?**]. In this work, we calculate the pairwise Spearman correlation between our extracted features and remove collinearity among these features. If the correlation between a pair of features is greater than a threshold, we only keep one of the two features in our model. In this work, we choose the correlation value of 0.8 as the threshold to remove collinear metrics, as suggested by prior work [**?**].

2.3 Modeling approaches

We formulate the task of predicting the receiver of a football pass as a learning to rank problem [**?**]. For each pass, our learning to rank model outputs a ranked list of the candidate receivers. A good model should rank the correct receiver in the front of the ranked list.

Gradient boosting decision tree (GBDT) is widely used for learning to rank tasks. There are quite a few effective implementations of GBDT, such as XGBoost and pGBRT, which usually achieves state-of-the-art performance in learning to rank tasks.

In this work, we use a recent implementation of GBDT, **LightGBM** [**?**], which speeds up the training time of conventional GBDT (e.g., XGBoost and pGBRT) by up to 20 times while achieving almost the same accuracy. We use an open source implementation of LightGBM that is contributed by Microsoft[3].

[[TODO: How we train, tune, and test]]

[[TODO: Please help refine the language]] We use a 10-fold cross-validation approach to train and test the model. For each fold, we hold out 10% of the data as test set. For the first fold, we continue to split the data into training and validation data, then we do a grid search to get the top three set of parameters according to the validation set performance. Then for each fold, we train three models with these three set of hyperparameters **[[TODO: list all the hyperparameters]]** on the training data. On test data, we use the three models to predict to get three sets of result, and then average the prediction results for each candidate, then use the averaged result to do ranking. We find that with model ensembling, the accuracies improve about 0.1%$\tilde{1}$.5%.

2.4 Evaluation approaches

We use ***top-N* accuracy** and **mean reciprocal rank (MRR)** to measure the performance of our models. *Top-N* accuracy measures the probability that the actual receiver of a pass appears in the top N predicted receivers (i.e., the N players with the highest predicted probability of being the receiver). *Top-1* accuracy measures the probability that the actual receiver of a pass is the player

[3] https://github.com/Microsoft/LightGBM

Table 1. A summary of players' passing statistics.

	Back-field	Middle-field	Front-field	Overall
Passing accuracy	86%	83%	79%	83%
Median passing distance (m)	17	14	11	14
Passing forwards ratio	74%	61%	50%	62%

Table 2. Five-number summary of players' passing distance.

Min.	1st Qu.	Median	3rd Qu.	Max.
0	9	14	20	70

with the highest predicted probability. Reciprocal rank is the inverse of rank of the correct receiver in an ordered list of candidate receiver (sorted by the predicted probability of being the receiver) [**?**]. The mean RR is the average of the reciprocal ranks over a sample of passes. The mean RR ranges from 0 to 1, the higher the better.

We use **10-fold cross-validation** to estimate the efficacy of our models. All the passes in the dataset are randomly partitioned into 10 sets of roughly equal size. One subset is used as testing set (i.e., the held-out set) and the other nine subsets are used as training set. We train our models using the training set and evaluate the performance of our models on the held-out set. The process repeats 10 times until all subsets are used as testing set once.

3 Results

This section discuss our answers to our research questions.

3.1 RQ1: what are players' passing patterns?

Overall, players' passing accuracy is 83%, and the passing accuracy decreases from the back field to the front field. Table 1 shows a summary of players' passing statistics. We define the passing accuracy as the ratio of the passes that pass the ball to a teammate. We divide the field into three equal-sized areas along the long side of the field, namely back field, middle field and front field. The passing accuracy for the back field, middle field, and front field is 86%, 83%, and 79%, respectively.

The median passing distance is 14 meters, and the passing distance decreases from the back field to the front field. Table 2 shows the five-number summary of players' passing distance. While the maximum passing distance is 70 meters, 50% of the passes are between 9 and 20 meters. The median passing distance for the back field, middle field, and front field is 17, 14, and 11 meters, respectively.

Players pass the ball forwards in 62% of the passes, and the ratio of forward-passing decreases from the back field to the front field. In

Table 3. The accuracy of our models for predicting the receiver of a pass (excluding false passes).

	Back-field	Middle-field	Front-field	Overall
Top-1 accuracy	53%	46%	53%	50%
Top-3 accuracy	84%	81%	90%	84%
Top-5 accuracy	93%	93%	97%	94%
MRR	0.70	0.66	0.72	0.68

Table 4. The accuracy of our models for predicting the receiver of a pass (considering all passes including passes to the other team).

	Back-field	Middle-field	Front-field	Overall
Top-1 accuracy	45%	38%	43%	41%
Top-3 accuracy	72%	68%	72%	70%
Top-5 accuracy	82%	80%	83%	81%
MRR	0.61	0.56	0.60	0.58

the back field, players pass the ball forwards in 74% of the passes, and the ratio decreases to 61% and 50% in the middle field and front field, respectively.

Players present different passing patterns in different areas of the field, which motivates us to extract features that capture the positions of the players in the field. Such different passing patterns also suggests us to construct and evaluate our models in different areas of the field separately.

3.2 RQ2: how well can we model the receiver of a pass?

Our models can predict the receiver of a pass with a top-1 accuracy of 50% and a top-5 accuracy of 94%, when we exclude false passes (i.e., passes to the other team). Table 3 shows the performance of our models in terms of *top-N* accuracy when we exclude false passes. The top-1, top-3, and top-5 accuracy is 50%, 84%, and 94%, respectively, which means the actual receiver of a pass has a 94% chance of being within our top-5 predicted candidates.

Our models can predict the receiver of a pass with a top-1 accuracy of 41% and a top-5 accuracy of 81%, when we consider all passes. Table 4 shows the performance of our models when we consider all passes (including false passes). The performance of our models decreases when we consider false passes (i.e., passes to the other team). False passes are very difficult to predict because it is not the sender player's intention to pass the ball to the other team.

Our models perform better when the sender of a pass is in the back and front areas of the field, while perform worse when the sender is in the middle area of the field. Table 3 and Table 4 also shows the performance of our models for the passes when the sender is in the back, middle and front areas of the field, separately. Surprisingly, the performance of our models is worst when the sender is in the middle area of the field. A player in the middle area may have more passing options, thereby increasing the difficulty to predict the right receivers.

Our models can predict the receiver of a pass with a top-1, top-3 and top-5 accuracy of 50%, 84%, and 94%, respectively, when we exclude false passes. Our models perform better when the sender of a pass is in the back or front area of the field.

3.3 RQ3: what are the influential factors that explain the passing targets?

4 Conclusions