Predicting Pass Receiver In Football Using Distance Based Features

Yann Dauxais[1] and Clément Gautrais[1]

[1] Univ. Rennes, INRIA, INSA, CNRS, IRISA
clement.gautrais@irisa.fr

Abstract. This paper presents our approach to the football pass prediction challenge of the Machine Learning and Data Mining for Sport Analytics workshop at ECML/PKDD 2018. Our solution uses distance based features to predict the receiver of a pass. We show that our model is able to improve prediction results obtained on a similar dataset. One particularity of our approach is the use of failed passes to improve the predictions.

Keywords: Pass prediction · Distance based features · Interception prediction.

1 Introduction

Thanks to the availability of football in-game events, recent approaches have studied aspects of the game that provide actionable knowledge for football clubs. For example, passing patterns from La Liga teams have been studied [8]. Other work focused on estimating football players performance from their actions on the field [1].

The football pass prediction challenge of the Machine Learning and Data Mining for Sport Analytics workshop at ECML/PKDD 2018 focuses on an aspect of the game that has been seldom studied. Indeed, few work have studied the problem of predicting the receiver of a pass in football. In [7], Vercruyssen et al. study the performance of predicting the receiver of a pass, based on different types of features: static or dynamic, quantitative or qualitative and relational or non-relational. They recommend using static qualitative features, such as coned based calculus [2] and double cross calculus [9]. This study is very relevant to our work, as the dataset they are studying is very similar to ours.

In [5], Steiner studies the effects of perceptual information on the probability for a player to receive a pass. He concludes that players with open passing lanes have a higher chance of receiving a pass. Players who are located forward of the passer, close to him or far from an opponent also have a higher chance of receiving a pass. This study is based on quantitative features, and shows pass prediction results similar to [7]. It should be noted that the data used in [5] is not based on real game data, but on scenarios that are evaluated by football players in an offline setting. This might have an influence on the pass decision process of the player.

Work have also been dedicated to the prediction of whether a pass will fail [3, 6], that is whether the opposing team will intercept the ball during the pass. Important factors are ball velocity, defenders distance to the ball carrier [6] and distances between players [3].

2 Dataset Presentation

The dataset $\mathcal{D}$ contains 12,124 passes performed during 14 different games involving a Belgian football club during the 2014/2015 football season[1]. For each pass, we have the x and y positions of all players, as well as the sender and receiver of the pass. We also have the number of seconds elapsed since the beginning of the half when the pass is performed and received. This yields a total of 60 columns ($2 * 28$ positions, 2 players id and 2 number of seconds). Players 1 to 14 belong to the home team; players 15 to 28 belong to the away team. A thorough description of the dataset can be found on the challenge repository[1].

Formally, we have $\mathcal{D} = \langle pass_1, \ldots, pass_j, \ldots, pass_{12124} \rangle$, with $pass_j = \langle x_1, y_1, \ldots, x_i, y_i, \ldots, x_{28}, y_{28}, i_s, i_r, t_s, t_r \rangle$ the j-th pass. x_i and y_i are the coordinates of player i, p_s is the pass sender id, p_r the pass receiver id, t_s the pass sending time and t_r the receiving time.

2.1 Dataset Exploration

We here briefly present some of the characteristics of the dataset. First, we have the positions of 28 players, but there are only 11 players on the field for each team. Fortunately, the substitutes are easily identified, as their x and y positions correspond to the Not a Number (NaN) token. Players having at least one coordinate equal to NaN are removed from the the pass vector $pass_j$.

This process yields a total of 22 x and y positions in $pass_j$: one for each player on the field. However, it can happen that a team has less than 11 players on the field, because of an injury or a red card for example. There are 438 pass examples (3.6% of the dataset) were less than 22 players are on the field.

In [7], the authors consider successful passes only: passes where the receiver and sender are on the same team. In the dataset, there are 2077 examples of failed passes (17.1% of the dataset). Because these passes represent a significant portion of the dataset, we decided to keep them.

We removed passes where the sender and receiver are the same players (6 cases), and passes were either the sender or the receiver have NaN positions (2 cases). The final dataset contains 12116 passes.

3 Model Description

The problem definition of the pass prediction challenge is quite simple.
Challenge problem: For each pass, given the sender and the positions of players, predict the receiver.

[1] https://github.com/JanVanHaaren/mlsa18-pass-prediction

3.1 General Approach

The idea of our approach is to estimate the probability P_i that the sender (player p_s) passes to player $i,\ i \neq p_s$. Then, sorting players id in decreasing order of probability P_i yields a ranking on the pass receiver prediction. Therefore, the player with rank 1 is the player with index equal to $arg\_max_{i\in[1,22],i\neq p_s}P_i$. We now present the learning of the probabilities P_i.

3.2 Learning pass probabilities

To learn the pass probabilities P_i, we train a classifier, that, given a set of features, outputs the probability that player i receives the pass from player p_s. Previous approaches predicting the receiver of a pass use different types of features. While the use of qualitative features (relative positioning of players, with respect to the sender and the receiver of the pass) is recommended [7], methods based on quantitative features (distances between players and pass lanes) yield similar results in the prediction of the pass receiver [5].

To train our classifiers, we use both quantitative and qualitative features. Given the position of the sender and of player i (the potential receiver), we first compute different distance based features. First, we compute the Euclidean distance between both players. Afterwards, we compute the forward distance of the pass, that the difference between player i $x$coordinate and the sender x coordinate. Then, we compute the smallest distance between an opponent of the sender and the pass lane. This corresponds to the smallest distance between the line formed by the sender and potential receiver and an opponent of the sender. Next, we compute the smallest and second smallest distance between the sender and players of the opposing team. We also compute the smallest and second smallest distance between the player i and players of the opposing team. Then, we compute a boolean indicating whether the potential receiver and the pass sender are in the same team. Finally, we encode the position of the sender and of player i using a regular grid over the football field. We split the field into 6 rows and 9 columns. Then, the position of a player corresponds to its grid number.

To summarize, we have 10 features to estimate the probability P_i that the sender passes to player i: the distance between the sender and player i; the forward distance of the pass; the smallest distance between the pass line and an opponent; the smallest and second smallest distance between the sender and an opponent; the smallest and second smallest distance between player i and an opponent; the sender and receiver grid number and whether the sender and player i are in the same team. With these 10 features, the classifier predicts whether this pass is the one chosen by the sender. If the sender chose to pass to player i, the value to predict is 1, and 0 otherwise.

4 Results

This section presents the pass receiver prediction results. This corresponds to evaluating the ranking yielded by ordering players id in descending value of P_i.

Pass type	MRR	top-1	top-2	top-3
All	0.886(±0.005)	0.841(±0.006)	0.890(±0.006)	0.915(±0.006)
Successful	0.909(±0.006)	0.861(±0.006)	0.919(±0.004)	0.947(±0.003)
Failed	0.779(±0.016)	0.746(±0.018)	0.752(±0.019)	0.763(±0.016)

Table 1. MRR and top-1, top-2 and top-3 recall for predicting the receiver id of a pass. The three lines correspond to the whole pass dataset, the successful pass dataset and the failed pass dataset respectively.

In the followings, this ordered list of players id is called V and v_j refers to the player id at the jth position in this list V. [2]

We train different classifiers (random forest, logistic regression and k nearest neighbors) to predict P_i. We found that the best performance was achieved when using a random forest with 200 trees. To evaluate the performance of the ranking, we use the same metrics as [7]: the mean reciprocal rank measure (MRR): $MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank}$, with n the number of examples, $rank$ the rank of the receiver ($rank \in [1, 21]$), and the recall in the top-1, top-2 and top-3 of the ranking. These recalls are equivalent to recall at k [4] for k set up to 1, 2 and 3 respectively: $R@k(V) = \frac{\sum_{j=1}^{k} rel(v_j)}{\sum_{j=1}^{m} rel(v_j)}$ where m is the length of V and $rel(v_j)$ is the relevance of v_j and $rel(v_j) = 1$ if v_j is the receiver id and 0 otherwise. It is worth noticing that one and only one player id is relevant for each pass what means that $\sum_{j=1}^{m} rel(v_j) = 1$ and $R@k(V) = \sum_{j=1}^{k} rel(v_j)$. This recall is then averaged on the pass set like for MRR: $\frac{1}{n}\sum_{i=1}^{n} R@k(V)$. Thereby, the top-1 recalls in the followings are equivalent to accuracy. Reported results correspond to the mean value of the corresponding metric on a 5-folds cross-validation. Table 1 presents the result for these 4 measures.

As one can see, the classifier predicts the good receiver from a list of 21 players in than 84.1% cases. This good classification result is improved to 91.5% when predicting the 3 most probable receivers. We can see by comparing the second and third rows of the table that it is easier to predict the receiver of a successful pass than of an intercepted pass. Furthermore, the prediction improvement between top-1 and top-3 for successful pass is 8.6% when the same improvement for intercepted pass is only 1.7%. It shows that the error on the receiver of an intercepted pass is much higher than for a successful pass.

To compare our results with the ones obtained in [7], we restrict ourselves to successful passes only, that is passes between two players of the same team. In this case, the $rank$ used for the MRR computation belongs to the interval $[1, 10]$.

[2] Our code is at `https://gitlab.inria.fr/cgautrai/prediction_mlsa2018.git`

Method	MRR	top-1	top-2	top-3
Best from [7]	0.42	0.279	0.416	0.467
Our Best (RF)	0.870(±0.006)	0.803(±0.009)	0.877(±0.005)	0.922(±0.005)
DT 555 leaves	0.664(±0.004)	0.507(±0.005)	0.675(±0.005)	0.777(±0.007)

Table 2. MRR and top-1, top-2 and top-3 recall for predicting the receiver id of a successful pass. The first line corresponds to the results obtained in [7] and the 2 other ones to ours: Random Forest (RF) and Decision Tree (DT).

The results are presented in Table 2. It is clear that our approach outperforms the results obtained in [7] on a similar dataset. Our MRR is 2 times better and our top-1 recall 3 times better. It should be noted that the best model in [7] uses 555 rules. To perform a more relevant comparison, we show the results for a decision tree having 555 leaves. This classifier still outperforms the results presented in [7], showing the relevance of our approach.

4.1 Learning from Failures

One of the differences between our approach and the one of Vercruyssen et al. [7] is that we consider opponents as potential receivers. While predicting who will intercept a failed pass is more difficult than predicting who will receive a successful pass, learning from both failed and successful pass can help correct some successful pass decisions. Indeed, when learning from successful passes only, one might over-estimate the probability of a dangerous pass to be successful. Using failed passes can help the classifier to identify these dangerous passes more accurately.

This effect can be observed from the analysis of Tables 1 and 2. Indeed, the MRR for predicting successful passes when using a classifier trained on both successful and failed passes is 0.909 (second line of Table 1); whereas the MRR for predicting successful passes when using the same classifier trained on successful passes only is 0.870 (second line of Table 2). One interesting thing to note is that the top-1 recall difference between both methods is of 5.8%, while the top-3 recall difference is of 2.5%. This means that while both approaches have a similar top 3 ranking for passes, training on both successful and failed passes leads to a better top 1 ranking. It is also worth noting that, for all passes, the MRR lower bound is equal to $\frac{1}{21}$ whereas, for successful passes only, this lower bound is $\frac{1}{10}$. Thereby, it is unfair to compare the MRR of all passes to the one of successful passes only.

4.2 Qualitative analysis

We now quickly analyze some predictions of our method by looking at 3 pass examples. These passes are represented in Figure 4.2.

The top-left example shows a case where our method correctly identifies the receiver: the label 1 is below the orange sign. The receivers ranked 2 and 3 are also teammates of the sender. The rank 2 player is unlikely to receive the ball, as he is close to 2 opponents. The goalkeeper, ranked 3, is however a safe pass option. We can see that our model suggests good passing options, even though the rank 2 player does not seem appropriate.

The top-right image also shows a case where our method correctly identifies the receiver. One interesting thing to note in that example is that the player ranked 2 is actually an opponent of the sender. An even more interesting fact is that this player is likely to intercept the pass between the sender and the actual receiver. This shows that our method is capable of putting high ranks for interceptions, if they make sense. Finally, the bottom example shows an case where our method is not able to predict the receiver in the top 3. While all options chosen by our method seem reasonable, the sender decided to choose another option. Overall, we can say that our method is able to output meaningful rankings, with some exceptions. These exceptions are mostly due to the fact that our method only takes into account distances. Adding new features can help to further eliminate unlikely passes.

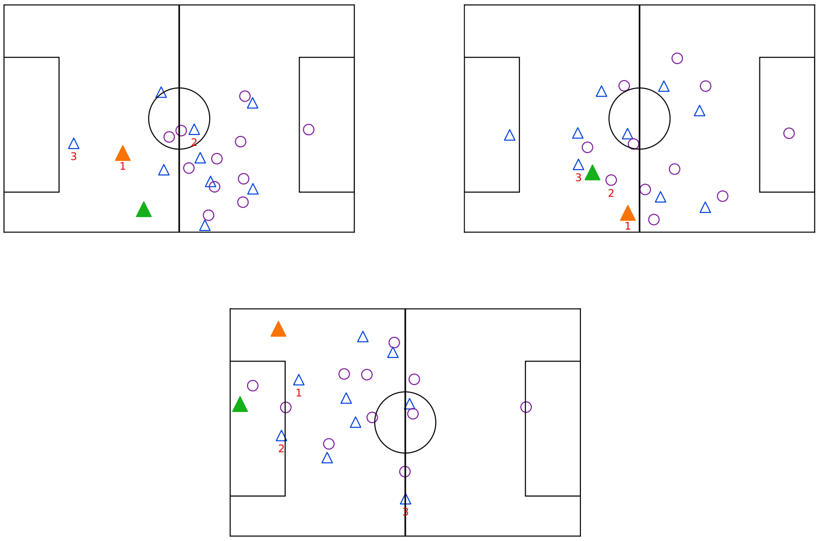

Fig. 1. Top 3 ranking for pass receivers in 3 cases. The pass sender is in green, the pass receiver in orange, the sender teammates in blue and the sender opponents in purple. The ranking is indicated as a red number below the symbol.

5 Conclusion

In this paper, we have illustrated how a simple method, using distance based features, is able to accurately identify the receiver of a football pass. This is done by computing the probability to receive the pass for each player and by predicting the potential receiver with the highest probability. Such approach allows to use a biggest feature set, an thereby, can be generalized. The particularity of this approach is that it also considers failed passes, and predicts the opponent intercepting the pass. For future work, we aim to add new features to strengthen the prediction of the receiver. A good starting point would be to add qualitative features described in [7].

References

1. Decroos, T., Van Haaren, J., Dzyuba, V., Davis, J.: Starss: A spatio-temporal action rating system for soccer. In: Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop (2017)
2. Frank, A.U.: Qualitative spatial reasoning with cardinal directions. In: 7. Österreichische Artificial-Intelligence-Tagung/Seventh Austrian Conference on Artificial Intelligence. pp. 157–167. Springer (1991)
3. Horton, M., Gudmundsson, J., Chawla, S., Estephan, J.: Automated classification of passing in football. In: PAKDD. pp. 319–330. Springer (2015)
4. McSherry, F., Najork, M.: Computing information retrieval performance measures efficiently in the presence of tied scores. In: European conference on information retrieval. pp. 414–421. Springer (2008)
5. Steiner, S.: Passing decisions in football: Introducing an empirical approach to estimating the effects of perceptual information and associative knowledge. Frontiers in psychology **9**, 361 (2018)
6. Travassos, B., Araújo, D., Davids, K., Esteves, P.T., Fernandes, O.: Improving passing actions in team sports by developing interpersonal interactions between players. International Journal of Sports Science & Coaching **7**(4), 677–688 (2012)
7. Vercruyssen, V., De Raedt, L., Davis, J.: Qualitative spatial reasoning for soccer pass prediction (2016)
8. Wang, Q., Zhu, H., Hu, W., Shen, Z., Yao, Y.: Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications. In: Proceedings of the 21th ACM SIGKDD ICKDM. pp. 2197–2206. ACM (2015)
9. Zimmermann, K., Freksa, C.: Qualitative spatial reasoning using orientation, distance, and path knowledge. Applied intelligence **6**(1), 49–58 (1996)