Valuing passes in football using ball event data

Lotte Bransen[1,2], Michel van de Velden (sup.)[2], and Jan Van Haaren (sup.)[1]

[1] SciSports
[2] Erasmus University Rotterdam

1 Introduction

Scouts at football clubs aim to discover players having a positive influence on the outcomes of matches. Since passes are the most frequently occurring on-the-ball events on the pitch, a natural way to achieve this objective is by identifying players who are effective in setting up chances. Unfortunately, traditional statistics fail to reveal players excelling in this area. For example, the percentage of successfully completed passes does not differentiate between a pass between two central defenders on their own half and a pass by a midfielder trying to reach a striker in the opponent's penalty area. However, while the latter pass is more likely to fail, it could also increase the team's scoring chances at the same time.

To overcome the limitation of traditional statistics, we propose two metrics to measure players' on-the-ball contributions from passes. The first metric computes the expected reward of each pass by analyzing similar passes in the past, whereas the second metric computes the *added* expected reward of each pass by analyzing similar patterns of play in the past. Due to the low-scoring nature of football, football players only get a few occasions to earn reward from their passes each match (i.e., each time a goal is scored). Therefore, our metrics compute the number of goals expected to arise from their passes if they were repeated many times. Both metrics operate on ball event data that describe all on-the-ball events that happen during a match. For each event, our dataset includes a timestamp, the team and player performing the event, its type (e.g., pass or shot), and its start and end location. Our dataset covers 9113 matches in the 2012/2013 through 2016/2017 seasons in five European top-tier leagues.

2 Approaches

We propose two novel approaches to assign values to passes based on their expected impact on the scoreline. Both approaches require splitting the event stream for each match into a set of possession sequences, which are sequences of events where the same team remains in possession of the ball. We label each possession sequence by computing its expected reward. When a possession sequence results into a shot, the sequence receives the expected-goals value of the shot, which reflects the probability of the shot yielding a goal based on the outcomes of similar shots in the past. When a possession sequence does *not* result into a shot, the sequence receives a value of zero.

The pass-oriented pass value (PPV) approach computes the expected reward of a pass by leveraging a k-nearest neighbors search with a domain-specific distance function, which considers both the characteristics of the pass and the circumstances under which the pass was performed. We value a pass by first performing a k-nearest neighbors search over all passes and then averaging the labels of the possession sequences of the k most-similar passes. Our distance function considers the characteristics of the passes as well as information derived from the sequences of events leading up to the passes, which provide valuable information about the circumstances under which the passes were performed.

The sequence-oriented pass value (SPV) approach computes the *added* expected reward of a pass by computing the difference between the expected reward of the possession subsequence after that pass and before that pass. Hence, a pass receives a positive value if the expected reward of the possession sequence after the pass is higher than the expected reward before the pass. To this end, we split each possession sequence into a set of possession subsequences. Each subsequence starts with the same event as the original possession sequence and ends after one of the passes in that sequence. We compute the expected reward of a possession subsequence by first performing a k-nearest-neighbors search and then averaging the labels of the k most-similar possession subsequences. We use the dynamic time warping (DTW) distance measure to determine the similarity between two possession subsequences. We first apply DTW to the x coordinates and y coordinates separately and then sum the differences in both dimensions.

3 Evaluation and results

We evaluate the predictive performance of our metrics by predicting the outcomes of future matches. We rate a player by first summing the values of his passes for a given period of time and then normalizing the obtained sum per 90 minutes of play. We use these player ratings to determine the expected number of goals scored by the teams in future matches. The PPV approach, which yields a logarithmic loss of 1.0377, outperforms the SPV approach (1.1893). In addition, we compare our approaches to a baseline approach that predicts the prior distribution over the possible match outcomes for each match. Our PPV approach also beats this baseline, which obtains a logarithmic loss of 1.0498.

Rating the passes in the 2016/2017 season using our PPV approach, we identify Cesc Fàbregas, Franck Ribéry, Mesut Özil, David Silva and Kevin De Bruyne as the players with the highest contribution per match.

4 Conclusion

In summary, we introduced two novel player performance metrics for football that measure players' contributions to creating goal-scoring chances by computing the expected rewards for their passes. Our experimental evaluation showed that our PPV approach outperforms our SPV approach as well as a baseline on predicting the outcomes of future matches.