Leiden University

Study programme Informatica & Economie

Analysing Possession Switches in Football using Subgroup Discovery

Name:	Roy de Winter
Date:	14/06/2016
1st supervisor:	Arno Knobbe
2nd supervisor:	Ricardo Cachucho

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Analysing Possession Switches in Football using Subgroup Discovery

Roy de Winter

Abstract

In this thesis, we analyse possession switches in football. The algorithmic approach used to analyse the possession switches is subgroup discovery. Subgroup discovery is an element of the research area of data mining. Studying possession switches is important because recapturing the ball usually cost a lot of effort for the defending team. Possession switches can occur in four different situations: due to a lost duel, an intercepted pass, the ball was kicked over the sideline or because a foul was made. To analyse these different situations, 62 different features (variables) are constructed. the features can be categorized in six different categories: distances, positions, counts, ratios, surfaces and possible passes. With these six categories, every important aspect of a football match is captured. These 62 features are recalculated for every frame (1/10th second) of nine matches played in the Eredivisie in the first half of the season 2015/2016. Out of the almost 600.000 frames more than half of the frames turn out to be unreliable. This is a pity because the outcome of subgroup discovery suffers under bad input. Based on the current data, subgroup discovery tells us that the most important variables which leads to a possession switch is the distance to the closest three players of the defending team.

Acknowledgements

I would like to thank Ruud van Elk for supplying the data, Arno Knobbe for supervising this research and Benjamin van der Burgh for offering me technical support.

Contents

Chapter 1

Introduction

This thesis is done to gain more insight into football and ball possession switches during football matches. The research is supervised by Arno Knobbe and is written as a part of the final project for the bachelor thesis of Informatica & Economie at the Leiden Institute of Advanced Computer Science (LIACS). The data used in this research is provided by Ruud van Elk and is collected during 9 Eredivisie matches.

1.1 Research area

The research area is sports analytics, where we focus on the team-based sport football, also known as soccer. Sports analytics does not differ too much from regular analytics besides that the focus is on sports instead of other applications. Analytics can be defined as: finding patterns in data and using such patterns to answer questions. Subgroup discovery [FF99,KZ02] is one of the algorithmic approaches to find those interesting patterns. Subgroup discovery is an element from the research area of data mining [WF05]. More about subgroup discovery and data mining can be found in Chapter 2. An important aspect in this research is translating knowledge about football into variables so that data mining techniques such as subgroup discovery can be applied. This translating knowledge into variables is also known as feature engineering [GE06]. The features translate situations which took place during a match in numeric attributes for every single situation.

1.2 Why is this research relevant?

In this thesis we analyse possession switches in football. This is done for two reasons: because recapturing ball possession usually cost a lot of effort for the defending team and because the most effective way to recapture ball possession is still unknown. Defensive teams wait on their own half and mostly expect the

opponent to make a mistake while offensive teams give a lot of forward pressure and try to recapture the ball right after they lost it their self. Both these strategies work to recapture the ball and we can conclude that the team wanting to score a goal should recapture the ball at some point. How this is done and which variables are the most relevant for recapturing ball possession is still a mystery. Do possession switches occur more frequently because the player in possession loses a duel, a pass was intercepted, the ball was kicked over the side line or because a foul was made? This research is an attempt at defining the cause of possession switches and the conclusion of this research can be used in football practice for preventing possession loss and for regaining ball possession.

1.3 What can be learned?

The data set contains a lot of training, positional and event-based data. A lot can be learned from this data but in this research, we only focus on the cause of a possession switch.

1.4 Research question

The research question therefore deals with ball recapture moments, possession gains, possession losses or as we will call them in this thesis: possession switches. The main research question is formulated the following way:

Using an algorithmic approach that analyses possession switches during football matches, which variables have the highest influence on these moments?

To answer the main research question, five other sub-questions are posed and answered in different chapters of this thesis:

1. What data is available for our research?
2. Which variables (features) can be extracted from the data?
3. What does subgroup discovery tell us about possession switches?
4. What do the most important subgroups found tell about the influence of the different variables?
5. Is the current quality of the data sufficient enough to answer the research question?

1.5 Earlier research

Football nowadays is analysed after every match by experts in the football area. Mostly these analyses by experts are based on their expert opinions and deal about several exciting moments during the match or about one player. The downside is that these analyses mostly deal about several moments or one match and not about multiple games over the whole team. This is where research and statistics comes in, for instance an article was written that deals with models able to predict the number of goals scored in one season [Cal16]. Other good analytics articles about football and sports can be found on the website Statsbom [Opt16]. A summation of 20 interesting articles from the last few years are represented in an article of Medium. [Wor16].

Not only matches and team performances are analysed, football players individually are studied as well. This is done because a lot of money is involved with the transfers of players in the worldwide football leagues [HCK16]. Football clubs want to make money out of their trained football players and do not want to overpay for new purchases. Therefore the economic valuation of individual players is important.

Basketball, rugby, handball, (ice)hockey and lacrosse have a lot of similarities with football and a lot can be learned from these sports. These similarities are based on the common fact that all the sports are team sports, players are free to move over the entire field and the main purpose of the sports is to score goals. Earlier research done to these similar sports is summed up in the following survey [GH16].

1.6 Thesis Overview

This chapter contains the introduction; Chapter 2 includes the definitions and preliminaries which give more insight into the aspects of data mining [WF05], subgroup discovery and feature engineering; Chapter 3 answers the fist sub-question about the available data; Chapter 4 discusses our feature engineering and the second sub question; Chapter 5 tells us how the targets are set and calculated; Chapter 6 deals with the experiments done and answers the fourth sub-question; Chapter 7 draws the conclusion which features and combinations of features are the most important for possession switches.

Chapter 2

Definitions

An introduction into data mining [WF05], feature engineering [GE06] and subgroup discovery [FF99,KZ02] is given, so the main goal of the research is clear and the reason why features are constructed can be understood.

2.1 Data mining

To understand the cause of possession switches, data mining [WF05] algorithms are used during this research. The algorithms can give insight in the domain, attributes and find regularities between the variables. Some of the variables are so-called *targets*, the targets in our case are the possession switches. The targets are the variables the data mining algorithms are trying to classify. Classification is a forecast of what will happen in new situations based on data from previous situations. This is often done by guessing the classification of new examples. In our case we are not only interested in the classification, but we are more interested in the understanding of where the classification came from. Which variables are able to classify the target and which values should they have? The variables in our case are the features. The features who have the highest predicting value are the most interesting features.

2.2 Feature construction

The features are constructed and based on football knowledge. From a football perspective, a possession switch can occur after four different events:

1. After winning a duel
2. After intercepting a pass

3. After kicking the ball over the side lines

4. After making a foul which is seen by the referee

The team in possession is during this research defined as the attacking team, the team out of possession is the defending team. Mostly these possession switches occur when the pressure on the attacking team is high. This pressure that leads to possession switches is translated into several features.

2.2.1 Features

Exactly how the features are constructed can be found in Chapter 4. The features are expressed in six different categories:

1. Distances.

2. Positions.

3. Counts.

4. Ratios.

5. Surfaces.

6. Possible passes.

Distances The distances are mainly the distance between the players from the opponent teams, distances between the players and the ball and distances between players and their direct opponent. During a match, all the players have a direct opponent. The direct opponents are the two players who have to cover each other. Usually this is a striker versus a defender, a midfielder versus another midfielder and a defender versus a striker. As can be expected, the direct opponent from the defender is the striker, but when the other team is in ball possession the striker has to cover the defender. This way the direct opponents are symmetrical.

Positions The positions are the positions of the players and the ball. When the position of the ball is in a 16 meter area, usually the pressure of the defending team is way higher compared to when the ball is on the half of the team in possession.

Counts The number of opponents close to the ball or player in possession is also important for pressure from football perspective. When it is crowded around the player in possession with opposing players, the player in possession is under much more pressure than when there are fewer players.

Ratios The ratio features tell us something more about relative closeness of the players in possession and the opponent. In football perspective, this could be more interesting than the counts because the counts only provide hard numbers whereas the ratios tell something about the defending players and opponent players at the same time.

Dominant regions Players have a region around them which is their dominant region because they can reach every point in the region faster than the opponents. The size of the dominant regions can be interesting. When the size is big, the player probably is under almost zero pressure, while when it is small the player is under a lot of pressure. Because at a certain time during the match, we do not know how tired the players are and what their reaction time will be, the assumption had to be made that: *All players have the same reaction time and top speed.*

Possible passes The last features tell something about the possible passes. An important factor in football is the number of players able to receive the ball at their current location. If the player in possession has a lot of teammates able to receive the ball, it is expected that the team is less likely to loose the ball.

2.3 Subgroup discovery

After all the features and targets are set, a subgroup discovery tool Cortana [DFK16, MK11] is used to do subgroup discovery [FF99, KZ02]. The subgroup discovery is preformed on a data set. Every row of the data set has their own calculated feature and target values. The goal of subgroup discovery is to find patterns in different rows that all identify a possession switch. This means that the subgroup discovery algorithm is searching for unique groups of rows where the target is equal to 1. The subgroups found are therefore a set of rows which represent one or more possession switches.

Every subgroup is checked for patterns, these patters can be seen as conditions that apply to these subgroups. A condition could for instance be: *The distance between the ball and closest 2 players of the defending team should be smaller or equal to 10 meters for a possession switch to occur*. Of course, not only the moments possession switches occur satisfy this condition but also moments where a possession switch did not happen. For instance, such a false alarm can occur when the ball is tried to be recaptured by the two closest defenders. The distance is smaller than ten meters so the condition is satisfied, but the player in possession passes the ball right on time to an other player, before the two defenders are able to recapture the ball. Such a false alarm is called a false positive while a true ball recapture moment satisfying the condition is a true positive. The true positives and

		In subgroup	
		True	False
Target	True	true positive	false negative
	False	false positive	false negative

Table 2.1: Confusion matrix.

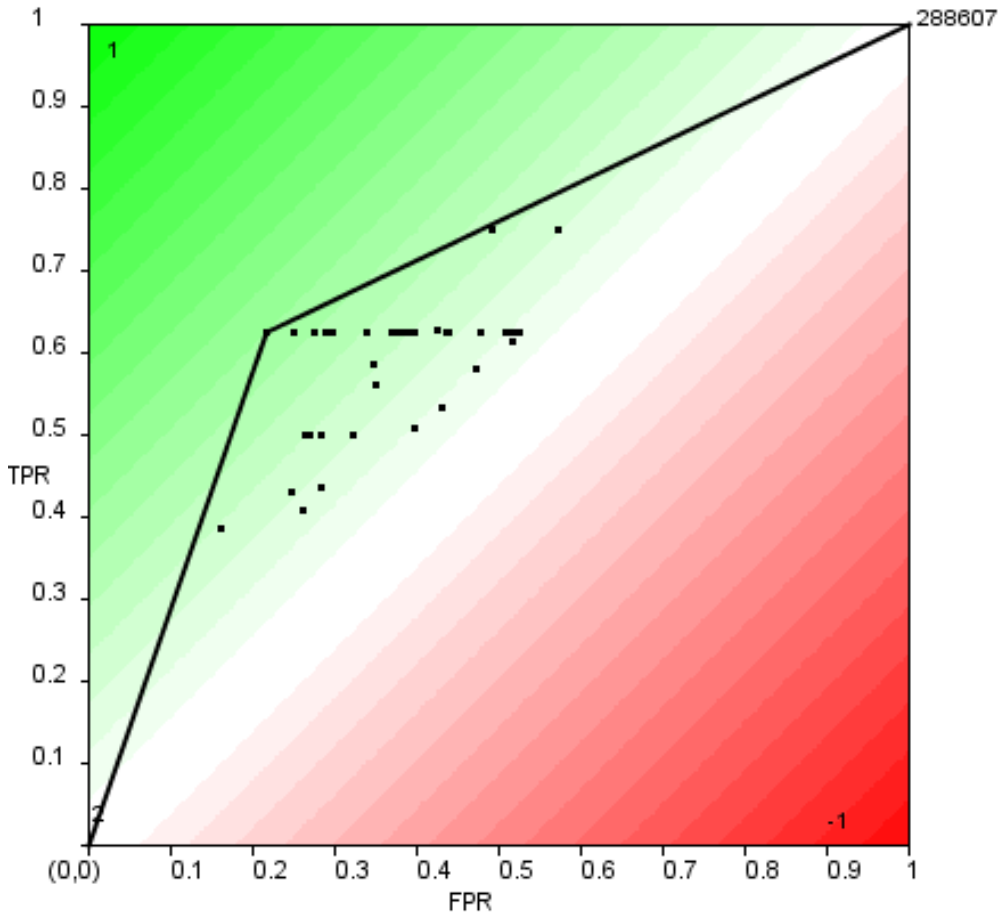

Figure 2.1: ROC curve in an ROC space.

false positives can be inserted in a confusion matrix represented in Table 2.1.

After all the subgroups are found and for every subgroup a confusion matrix was made, every subgroup can be placed in the Receiver Operation Characteristics (ROC) space. Figure 2.1 is an example of such an ROC space. The ROC space has on the horizontal axis the false positive rate and on the vertical axis the true positive rates. The interesting subgroups are the subgroups that have the highest performance. A high performance can be accomplished with a high true positive rate and a relatively low false positive rate. The most interesting subgroups can be found on the ROC curve which is formed by a line from (0,0) to the highest performing subgroups to (1,1). The area under the curve (AUC) is the overall performance [MG02] of the multiple subgroups in the data set.

Chapter 3

Data description and football domain

In football, a lot of data is gathered during matches and training sessions. During training sessions, players wear sensors which provide the trainer for instance with heartbeat, position, speed and acceleration of players. Not only during training sessions, data is gathered but also most of the professional football matches are recorded. Different parties gather data-based on these recordings. This data is already used to calculate the percentage of ball possession per team, shots on target, passes per team, location of players etc.

The data is available for this research contains training data, positional data and event data. In this thesis, we leave the training data out of scope and concentrate only on the positional and event data. The data is measured and gathered during nine matches in the Eredivisie. The matches took place in the first half of the season of 2015/2016.

3.1 Positional data

An important aspect of football is where the players and the ball are located on the field at a certain time during a match. The location of the players determines for instance whether a player can receive the ball, if he stands offside, if he possesses an attacking position, if he is in the position to make a goal attempt or if a foul can be made without causing a penalty for the opponent. Logically this is why the positions of the players and the ball are tracked during matches.

The tracking is done by three cameras with different angles which register the whole match. The camera systems processes ten frames every second which results in ten x- and y-coordinates every second, for each

player, for each referee and for the ball. Every 1/10th of a second is what we will name a frame from now on.

Table 3.1 represents five objects on the field in the 40th second from the first match. The coordinates represent the distance away from the center spot of the field like visualised in Figure 3.1. As you can see in Table 3.1, Visiting Team Player 2 has the x and y value of 99.999. This means that at this time, he is not registered by the camera system.

Timestamp (ms)	X	Y	Name
40,000	17.027	5.364	Home Team Player
40,000	-4.051	-3.378	Visiting Team Player 1
40,000	99,999	99,999	Visiting Team Player 2
40,000	11.094	-10.509	Referee
40,000	38.09	-7.985	Ball

Table 3.1: Frame 40,000 (first frame of 40th second).

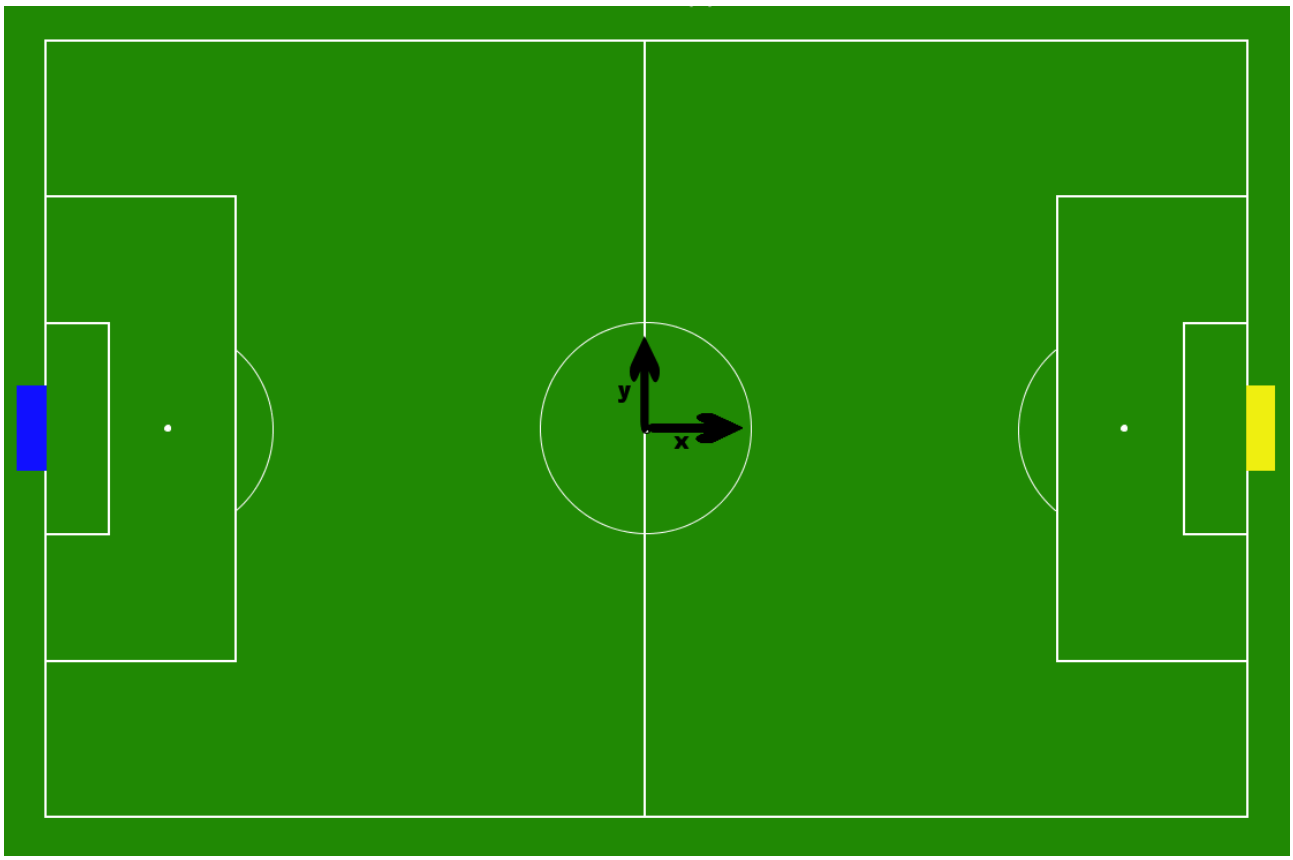

Figure 3.1: Football field x y.

Because the system only uses three cameras from one side of the field it sometimes happens that the system fails to track the ball or players who are out of sight. In this case, the system saves the values 99,999 for the coordinates unknown. The system has the most trouble tracking the ball, the ball coordinates are missing in 164,938 out of 592,194 frames which is 27.85%. Players are easier to track because they move not as fast as

the ball and they are relatively big compared to the ball. This is why relatively fewer player coordinates are missing. In total 1,923,503 out of 13,028,268 coordinates, 14.76% is missing for the player measurements. One cause of so many missing coordinates is the break, the break is also recorded by the cameras.

There are two more problems that occur while tracking the ball. The first one is that there is no z-coordinate registered. This first problem leads to the problem that it is hard to tell if the ball is in the air or not. Therefore the assumption: *The ball is always on the ground* had to be made. The second problem is that the camera is not positioned above the field but aside of the field. A high ball therefore sometimes looks like it is on the ground several meters further away than its actual position at the y-coordinate like visualized in Figure 3.2.

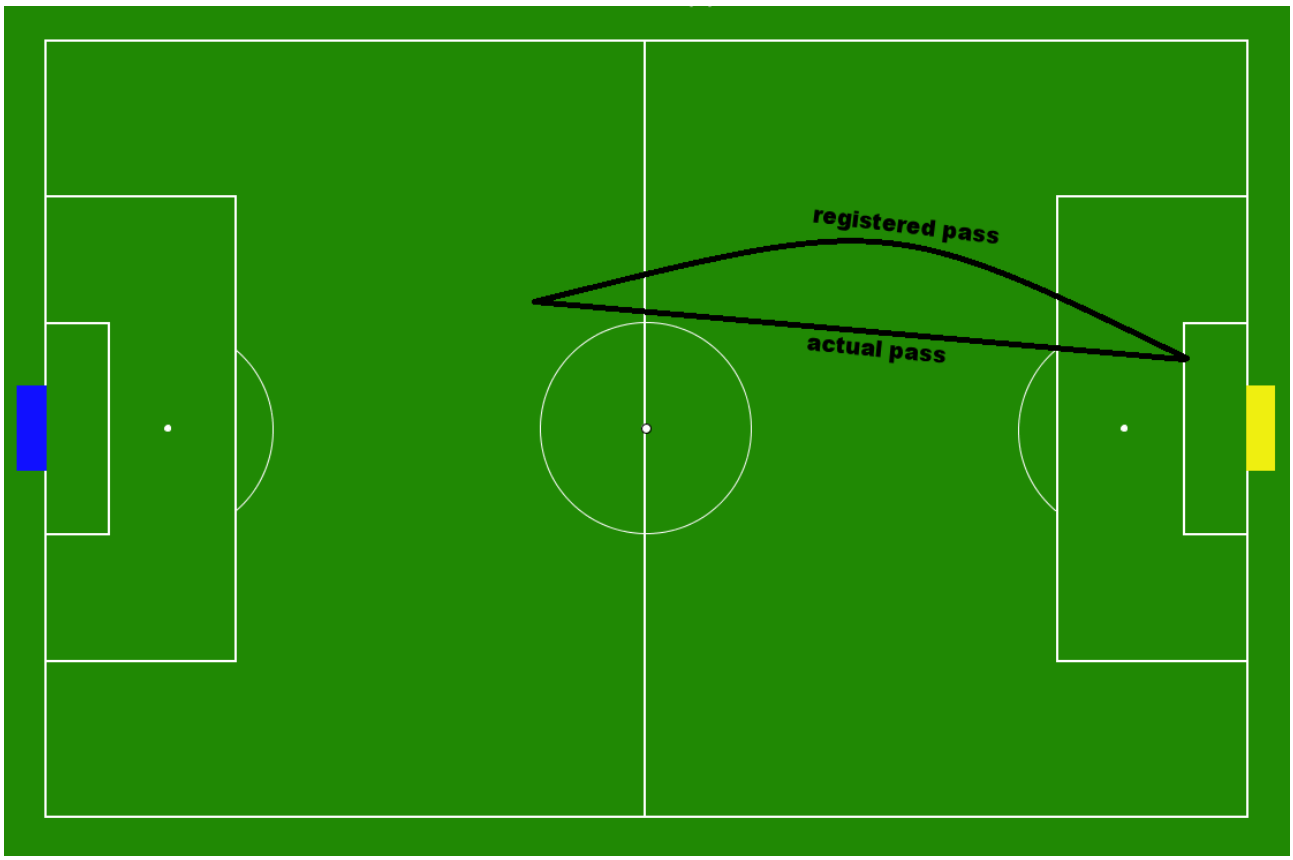

Figure 3.2: False high pass registration (bird his eye point of view).

When visualizing the data, the data turned out to contain an other element of noise. In some frames the camera system mixes players up. This is not that big of a problem if it happens between players from the same team but when a player from team 1 and team 2 are mixed up this could be problematic.

3.2 Event data

Another important aspect of football is the events occurring during the football matches. A short pass done with the right foot of Visiting Team Player 9 from Visiting Team is an example of an event. But also events like a throw in, dribble, attacking action, defending action, goal attempt, save on goal, goal, foul and indirect free kick are examples of events. When reading the events sequentially you can actually understand quite well what happened in the particular match the event data is about.

All these events are analyzed with a computer operated by people who are watching the game. The operators save information about when the events happen in mini seconds specific, in which half, the category of the event, the player in action, the team in action, the body part used, the definition of the action and of course which match it happened in. Five rows example of event data can be found in Table 3.2. The definition we are looking for is the *isPossessionLoss* and the *isPossessionGain* because this can tells us that the ball changes possession.

Time(ms)	Half	Category	Player	Team	Attribute	Definition	Match
7,132	1	attacking action	Player 8	Home Team	head — duel touched	isPossessionLoss — isDuelPart — isDuelAir — isAerial	Match 1
10,905	1	pass	Player 15	Visiting Team	right foot	isPassCompleted — isPassWide	Match 1
11,930	1	pass	Player 16	Visiting Team	right foot — direct	isPossessionLoss — isPassForward — isPassLong	Match 1
14,931	1	pass	Player 3	Home Team	direct — left foot	isPossessionGain — isPossessionGainInterception — isPassCompleted — isPassForward	Match 1
17,030	1	attacking action	Player 4	Home Team	duel touched	isDuelPart — isDuelWonByAttacker — isDuelStanding	Match 1

Table 3.2: Event data example.

Before we can conclude that the event data is correct, we have to keep in mind that event data is in some cases objective. For instance, a pass can be interpreted multiple ways, someone could say it is just an ordinary pass forward to one of his teammates, while others say it is a through pass. An other example of objectivity could be a cross pass that ends up in the goal. Is this pass a cross pass or a goal attempt? The last thing to keep in mind is that the operators of the computer are human with a limited reaction time. Because of this, not every event is rightly timed.

3.3 Data combined

The positional data is stored into a normalized MySQL database. An enhanced entity relationship model of the database is represented in Figure 3.3. In the database, the x and y coordinates are converted to floats between 1 and -1. This is done so the data can be easily visualized. The event data is saved in the same

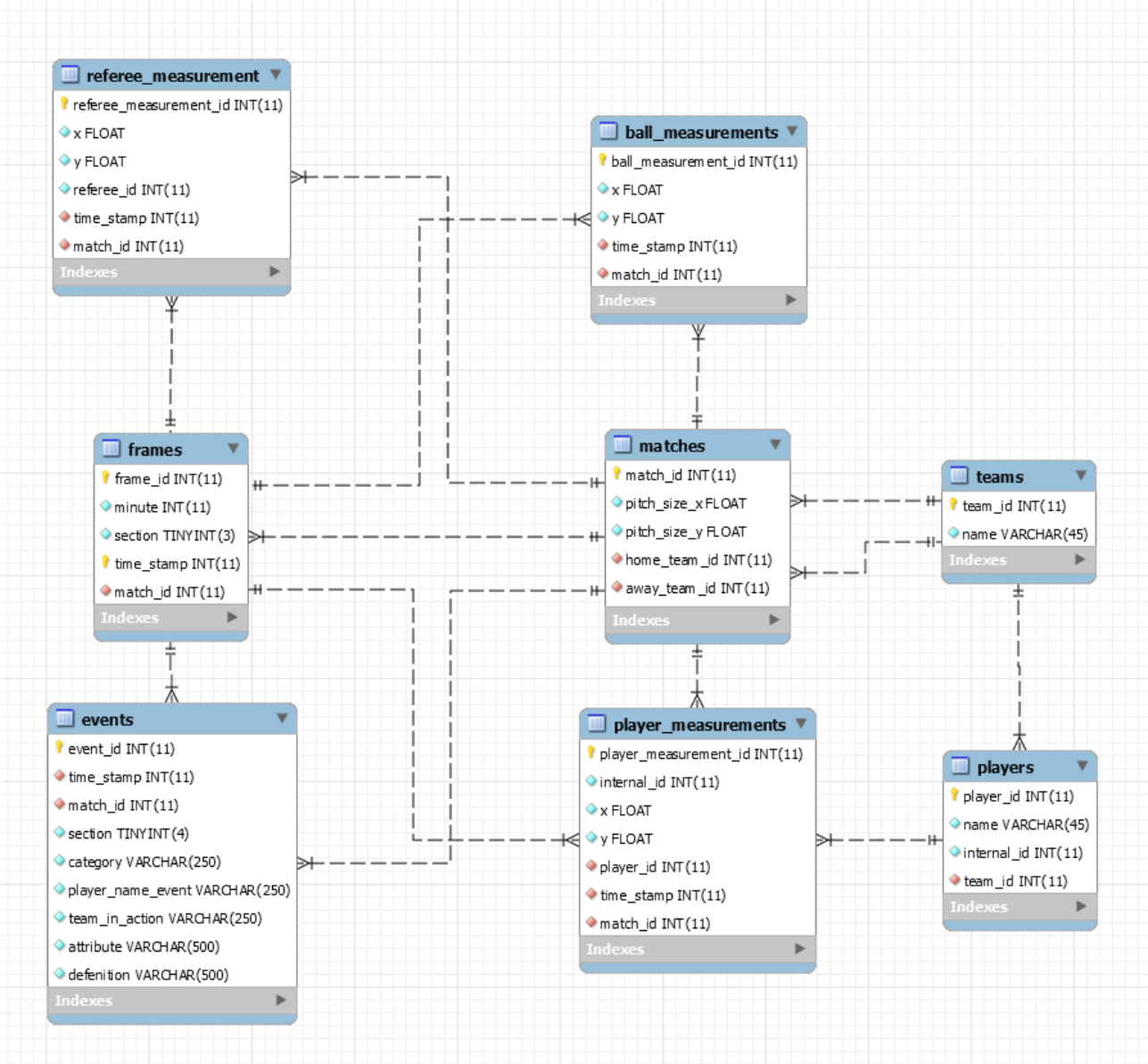

Figure 3.3: Enhanced entity relationship model representing the database.

MySQL database. Because the positional data only registers 1/10th of seconds, the event data is rounded to the closest 1/10th second so that their timestamps can be linked. After this process, the data can be used to visualize matches, features can be extracted and targets can be set.

3.4 Fill gaps in data

Some gaps in the positional data are really small and can be easily filled with logical values. This is done for gaps smaller or equal to five frames. This is done for missing x positions and missing y positions with Formula 3.1. This formula turns out to be pretty effective because 42.31% of the player gaps and 25.13% of the ball gaps could be filled. In total this formula filled 2,395 out of 5,661 player gaps, 7,559 out of 1,923,503

missing player coordinates, 367 out of 1,521 ball gaps and 1,254 out of 164,938 missing ball coordinates. The gaps larger than five frames are not filled, otherwise player and ball movements would be created, which might not even have taken place during the actual match. Still a reasonably large number of coordinates are missing because the breaks are also recorded.

$$\begin{aligned} missingValues &= FrameNumberEndOfGap - FrameNumberStartOfGap \\ lostPosition &= lastValueBeforeGap \\ foundPosition &= firstValueAfterGap \\ gapsize &= |lostPosition - foundPosition| \\ stepSize &= gapsize/(missingValues + 1) \\ \forall missingValue \in missingValues, missingValue &= lost + missingvalue * stepSize; \end{aligned} \tag{3.1}$$

3.5 Ball possession

Ball possession is not completely provided by the data. Possession can be seen in two different ways, the team in possession and the player in possession. Before the player in possession can be set, the team in possession should be known. Therefore for every frame the team possession is calculated first before the player in possession is set.

3.5.1 Team possession

The team in possession is an important aspect because a possession switch is what we are interested in and what is important for a lot of features. The possession is calculated in three different steps: set possession derived from event data, correct for closest player and last but not least correct for possession length.

Set possession derived from event data

For setting the team in possession, the event data is used. The event data proves the team that made an action at which frame. Some events that take place have an attribute that says *untouched*, these frames are not included. Not in every single frame an action was made, therefore the frames missing are filled with the possession from the frame before, this way roughly for every frame the possession was set.

Correct for the closest player

Because the event data is not precisely timed the possession changes do not completely match with what happens according to the positional data. Therefore the timing of the possession switch is incorrect and should be fixed. This is done by checking which player touched the ball for the last time. The player who touched the ball for the last time is not sufficiently enough supplied by the event data and therefore the player who touched the ball for the last time is set by checking the positional data. The player who touched the ball for the last time extracted from the positional data is the player who was closer or equal then two meters of the ball. This having been set for every frame the timing of the ball possession switch and therefore the ball possession per frame can be easily adjusted. Four cases can occur:

1. A possession switch is upcoming and the event data tells team 1 is in possession but a player from team 1 has not touched the ball yet.

2. A possession switch is upcoming and the event data tells team 2 is in possession but a player from team 1 already touched the ball.

3. A possession switch is upcoming and the event data tells team 2 is in possession but a player from team 2 has not touched the ball yet.

4. A possession switch is upcoming and the event data tells team 1 is in possession but a player from team 2 already touched the ball.

In case 1 and 3, the possession was adjusted so the possession switch took place later, right at the moment the ball was touched by the player playing for the team in the new possession. Figure 3.4 represents such a timing correction. For case 2 and 4, the possession was adjusted so the possession switch took place earlier, right at the moment the ball touched by the new team.

Correct for possession length

Possession for a team means that the ball is circulating in the team and that the opponents try to recapture the ball. Sometimes it happens that an opponent glances the ball but which not lead to a possession gain. When this happens the opponent team did make an action according to the event data, the ball was touched by the player from the opponent team and therefore a possession switch occurred according previous calculations. Because there was no possession switch this should be corrected. This is done by checking the length of the possession, if a team is in possession for less than or equal to two and a half seconds the possession is toggled to the team passing the ball around. Figure 3.4 represents such a correction for a glance.

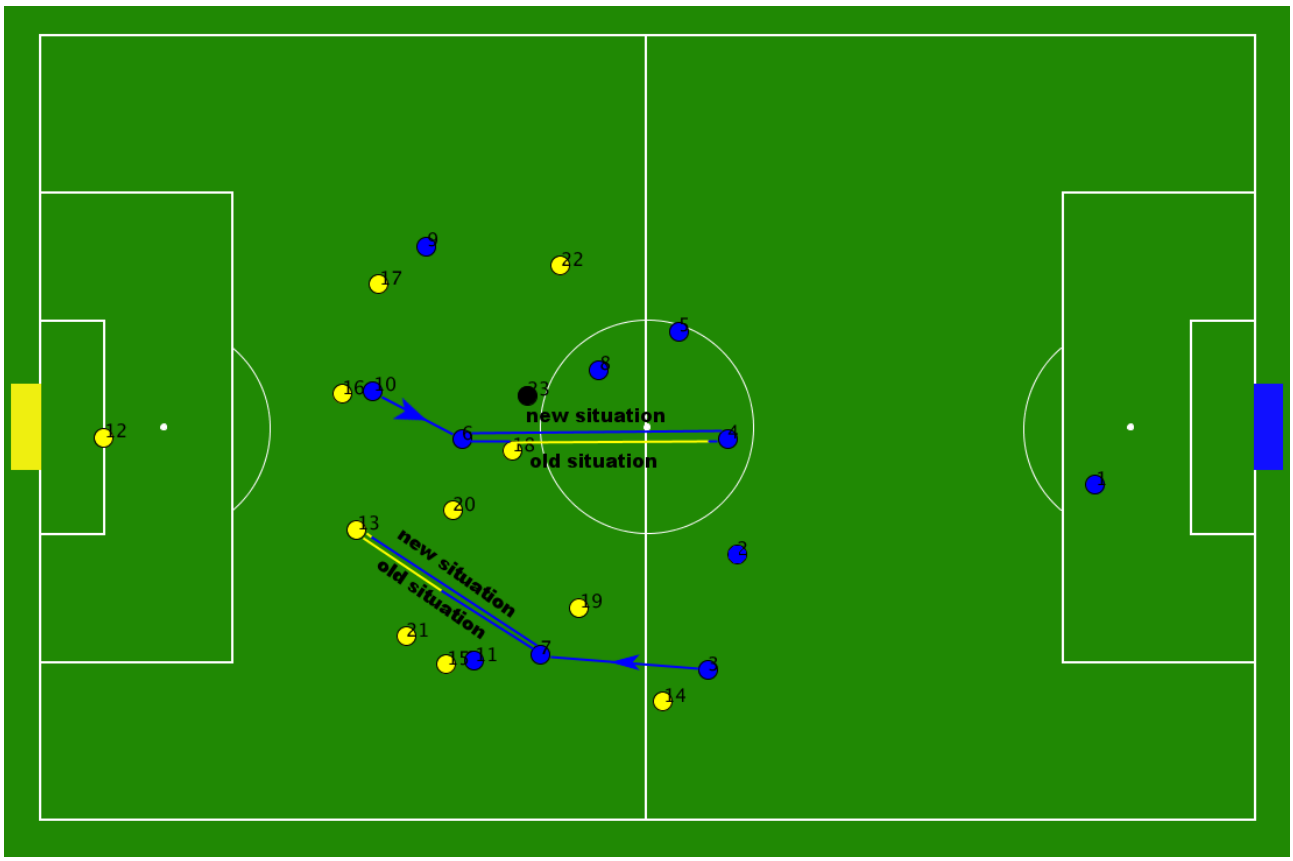

Figure 3.4: Team possession correction for closest player and for possession length.

3.5.2 Player possession

Now the possession per team is set the player in possession can easily be calculated. This is again done by checking which player was for the last time closer than or equal to two meters of the ball. This player of course should be a player who plays for the team currently in ball possession. This way during a pass from player 1 to player 2 the player in possession stays player 1 until player 2 is closer than or equal to two meters of the ball.

Chapter 4

Features

The data in the database now only provides us with a time stamp, the locations of the ball and the location of the players. This alone can not tell anything about possessions switches, that is why several features are made and recalculated for every single frame. The goal of these features is to translate football knowledge into variables that can predict possession switches. One problem with the features extraction is that the number of features and combination of features are is limitless and time is not. Therefore we tried to capture the most important features during this research. The features are separated and explained in the following sections: time stamps, distance, positional features, counts, ratios, dominant surfaces and possible passes.

4.1 Time stamps

The time stamps are the first two features. The first feature is the time stamp over the whole match. This feature is added because it can give an indication on how tired the players are and how much time the players have left till the end of the match or break. Because a player can regain its strengths and rest during the break, the second feature is the time stamp from the beginning of the second half.

4.2 Distance

This section deals with distances, in specific the distances between the ball and players, the distances between the entire team in possession and the opponent team and the distances between players and their opponents. All these features are calculated in meters using the euclidean distance metric [UC08]. If one value is missing the distance will become the maximum distance in the football field. Formula 4.1 is therefore the distance function used during this research. Missing values can occur if a player or ball is not registered by the camera.

$$distance(x_1, x_2, y_1, y_2) = \begin{cases} \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} & \text{if no missing values} \\ \sqrt{105^2 + 68^2} & \text{if missing values} \end{cases} \tag{4.1}$$

4.2.1 Ball

The following fifteen features try to capture pressure on the ball. The fifteen features are expressed in meters between the ball and certain players. All the features are calculated using Formula 4.2.

$$meters(players, pLocations, bLocation) = \sum_{p=1}^{players} Distance(pLocation, bLocation) \tag{4.2}$$

Ball distance to attacking team (1 feature)

The ball distance to attacking team is the sum of the distances between the attacking players and the ball. This is represented by the cumulative length of the yellow lines in Figure 4.1 given that the yellow team touched the ball for the last time. This can give an indication of the support the player in possession gets from his team mates. The total distance can be calculated by filling in Formula 4.2 with 11 for *players*, the locations of the attacking players and the location of the ball.

Ball distance to defending team (1 feature)

The ball distance to defending team is the sum of the distances between the defending players and the ball. This is represented by the cumulative length of the blue lines in Figure 4.1 given that the yellow team touched the ball for the last time. This can give an indication of the pressure on the ball from the defending team. The total distance can be calculated by filling in Formula 4.2 with 11 for *players* and the locations of the defending players and the location of the ball.

Ball distance to p closest players attacking team (6 features)

The ball distance to the p closest players from the attacking team can be calculated by choosing each of the following values 1, 2, 3, 4, 5 and 6 for p. First the distance between the ball and every team mate is calculated and put in a list. The list is ordered and the p smallest values are summed. The outcome of the summation indicates the support the player in possession gets from the closest p players. This is again calculated with the Formula 4.2.

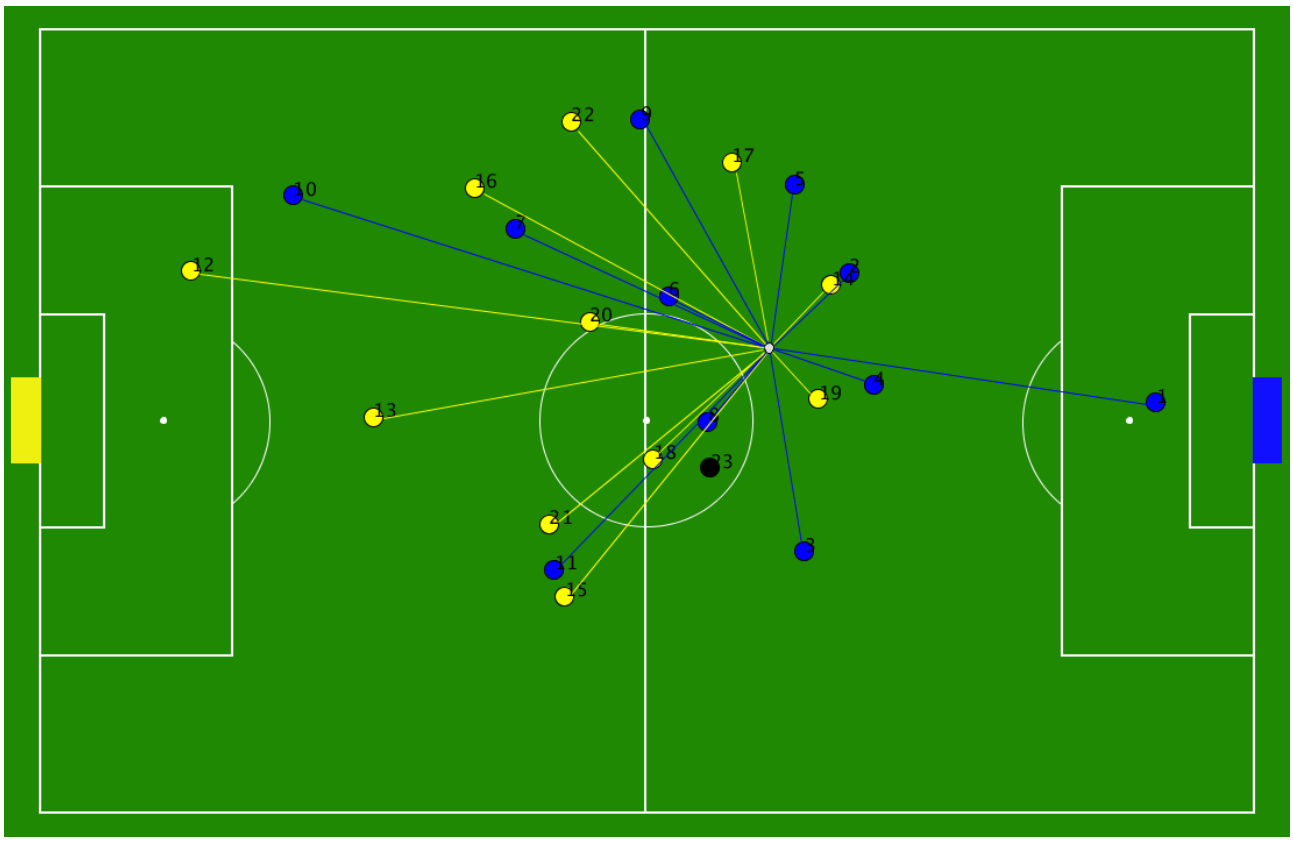

Figure 4.1: Distance to the ball.

Ball distance to p closest players defending team (6 features)

The ball distance to the p closest players defending team can be calculated by choosing a value from the list 1,2,3,4,5 and 6 are chosen for p. These indicate the pressure on the ball from the closest p opponents. First the distance between the ball and every opponent is calculated and put in a list. The list is ordered and the p smallest values are summed. This again can be calculated with the Formula 4.2

Ball distance between player in possession and the ball (1 feature)

The distance between the player in possession and the ball indicates if the player in possession has the ball close to him or made a pass to somewhere else. The feature can easily be calculated by filling in the Formula 4.2 with the values 1 for *players*, location of player in possession and location of the ball. Since the player in possession is the player who touched the ball for the last time it is not necessary true that this is equal to the distance between the one closest player of the attacking team.

4.2.2 Players

The following features tells us something about the pressure on the team in ball possession and the player in possession. The features can be calculated with the Formula 4.3 or the Formula4.4. What is good to know is that the keepers are not considered in this section and that the distances between teams are symmetric. This means the distance between team 1 and team 2 is equal to the distance between team 2 and team 1.

$$meters(pLocations, oLocations) = \sum_{p=1}^{10} \sum_{o=1}^{10} Distance(pLocation, oLocation) \tag{4.3}$$

$$meters(players, pLocations, oLocations) = \sum_{p=1}^{players} Distance(pLocation, oLocation) \tag{4.4}$$

An important fact is that every player has one direct unique opponent. This means that every player from team 1 has a direct opponent in team 2 which is his and only his direct opponent at that moment. The direct opponent in this research is stated as the optimal direct opponent for the whole team. So every player has one optimal direct opponent. This optimal direct opponent can be seen as an assignment problem which can be solved using a hungarian algorithm [Kuh55]. The parameters for the hungarian algorithm are the distances between the players from team 1 and team 2. This way the distances between the players and their direct opponents for the whole team is minimized. Figure 4.2 represents the optimal assignment for a certain frame during the first match.

Players total distance (1 feature)

Players total distance is the distance between team 1 and team 2. As explained in the paragraph above, this is a symmetrical problem. The connections between the players and opponents can be seen as a complete bipartite graph with ten nodes representing the player from team 1 and ten nodes representing the player from team 2. Weights on the edges could represent the distance between player $x_{1...10}$ and opponent $y_{1...10}$. The total distance between the two teams can be calculated with the Formula 4.3 with the players from team 1 and the players from team 2 as the arguments.

Direct opponent distance (1 feature)

Every player has one unique direct opponent. The direct opponent can be recalculated for every frame. The connections between the players and their unique opponent can mathematically be seen as a complete

matching bipartite graph. The graph can be extended with weights on the edges which represent the distance between a player from team 1 and his direct opponent from team 2. The distance between the players from team 1 and their opponent in team 1 is again symmetrical and represented in Figure 4.2. Formula 4.4 is used to calculate the summation of the distances. The parameters used are logically 10 for *players*, the players locations and their direct opponent locations. In football terms, this feature is important because it gives an indication of how close the direct opponent are to each other which could be important for pressure measurement.

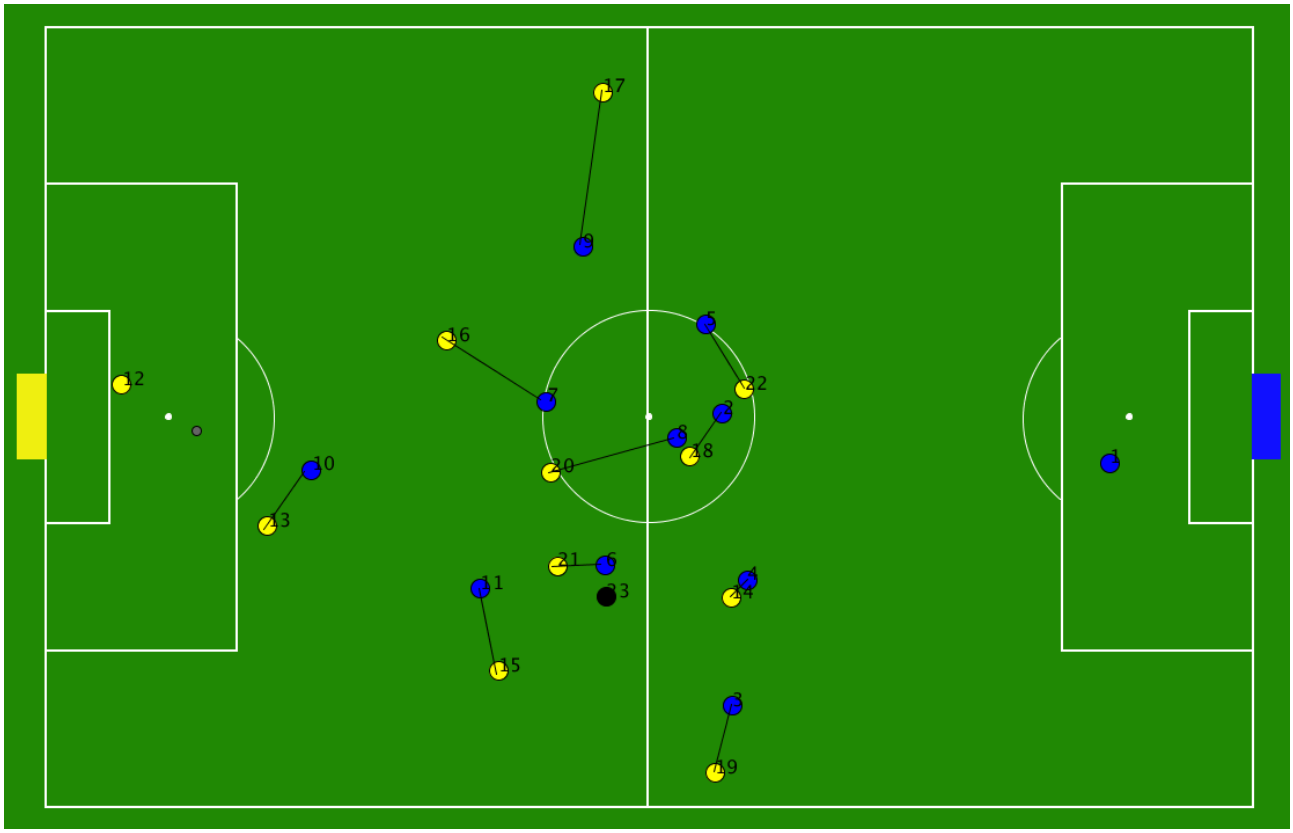

Figure 4.2: Optimal direct opponent.

Direct opponent whole match distance (1 feature)

The direct opponent for every player can also be calculated after the whole match. This means that every player has one unique opponent during the whole match. This can again mathematically be represented as a complete matching bipartite graph and can be seen as a symmetric problem. The distance can be calculated with the Formula 4.4 using the parameters 10 for *players*, players location from team 1 and their direct opponents location over the whole match from team 2. In football terms, this feature is important because it gives an indication of how close the team is to their direct opponent over the whole match.

Player in possession distance to the p closest opponents (6 features)

The distance between player in possession and the p closest opponents is calculated by calculating the distance between the player in possession and all the opponents. The values are inserted in a list which is ordered in an ascending way. For every p element of the list 1,2,3,4,5 and 6 a feature is made summing up the first p values of the ordered list. This feature is important from a football perspective. The cumulative distances between the player in possession and the p closest players can give an indication of under how much pressure the player is.

4.3 Positional features

This section deals with two positional features. The positional data give an indication of who is losing the ball and where the ball possession switch took place.

Ball distance to opponent goal (1 feature)

The ball distance to the opponent his goal can be calculated by checking which team is in ball possession and calculating the distance between the ball and the opponent his goal. This gives an indication of where the possession switch took place. This could be an important feature from a football perspective because losing the ball on your own half is usually more dangerous than losing the ball on the opposing half.

player in possession position average over match (1 feature)

The position for each player is expressed in the average distance between the back line and the player. So for an attacking player the average distance is bigger than for a defending player. This way every player gets a unique possession which is an indication of how offensive the player is. It is expected that players who are very offensive loose the ball more often than players who are really defensive. This is because usually the defensive players recapture ball possession and offensive players are the ones losing it.

4.4 Counts

The counts in this section are a count of players. A count of defensive players and attacking players. The counts can tell us something about the distribution of the players on the field.

Defending player count m meters (4 features)

For this feature, defending players close to the player in possession are counted. The players in between m meters are counted where m can be 5, 10, 15 or 20 meters. The counts limited by the distance of m can tell something about the amount of players close to the player in possession and thus the pressure on the player in possession.

Free offensive player count m meters (6 features)

For this feature, every free player from the team in possession is counted. A free player for this feature is defined in six different ways depending on the value of m. m can therefore take six different values: 2, 3, 4, 5, 6 and 7 meters. The count is done by checking how many players are more than m meters separated from their closest opponent.

4.5 Ratios features

The ratio features tell us something more about relative closeness of the players in possession and the opponent. These ratios are expected to be important in combination with other features.

Possession team versus opponent team, m meters (4 features)

This ratio gives an indication of the surroundings of the player in possession. This is done by dividing the count of teammates by the count of opposing players in between m meters. Four different values are chosen for m: 5, 10, 15 and 20 meters. If no opposing players in between the m meters are found then the ratio is set to the sum of team players plus one. The expected result of this feature is that when the ratio is low, the player in possession is more likely to lose the ball compared to when the ratio is high. The downside of this feature is that there is no information on which players of which team is closer

Possession team distance versus opponent team distance, p competitors (6 features)

This feature is calculated by dividing the two distances from the player in possession and his nearest p team players and the distance from the player in possession and his nearest p opponents. For p different values are chosen: 1, 2, 3, 4, 5 and 6 which will result in six different features. For instance, if p is 2, the distance between the player in possession and the two closest team players are calculated, the distance between the player in possession and the two closest opposing players are calculated and finally these two distances are

divided. The ratio this feature delivers can tell us something about the support from the teammates versus the opponents and the relative closeness between them and the player in possession. The downside of this feature is that there is no absolute information on how close the players are.

4.6 Dominant Surfaces

This section deals with dominant regions from different teams. The dominant regions are calculated with voronoi cells [PS85]. The voronoi cells are calculated using a voronoi diagram builder from vivid solutions [Dav16]. Remember that the assumption had to be made that every player has the same reaction time and top speed and every player has their own coordinate. Eleven features are extracted from these voronoi cells.

Voronoi surface team (2 features)

This feature tells something about the field possession and the sizes of the dominant region per team. For every player with their unique coordinate, a voronoi cell was calculated. The surface of the voronoi cell is calculated with the Formula 4.5 with x_n and y_n as the voronoi corner coordinates as the parameters. The surfaces of the teams are summed separately and the Voronoi surface of team in possession and the voronoi surface of the opponent team become two features. The surfaces give an indication of the dominance of the field for both of the teams. Figure 4.3 represents the problem where blue is the team in ball possession.

$$surface(x_{1\ldots n}, y_{1\ldots n}) = \frac{(x_1 \cdot y_2 - y_1 \cdot x_2) + (x_2 \cdot y_3 - y_2 \cdot x_3) + \cdots + (x_n \cdot y_1 - y_n \cdot x_1)}{2} \tag{4.5}$$

Voronoi surface ball (1 feature)

Because the ball is not an actor in the field the ball does not have its own voronoi surface. Of course the ball does have an x and a y coordinate. This coordinate must be in one voronoi cell belonging to one of the field players like represented in Figure 4.4. The surface of the voronoi cell belonging to the player is saved as a feature. This feature tells something about the surface of the dominant region of the player closest to the ball which can be important in football terms because this player could be the ball receiver or intercepting a pass.

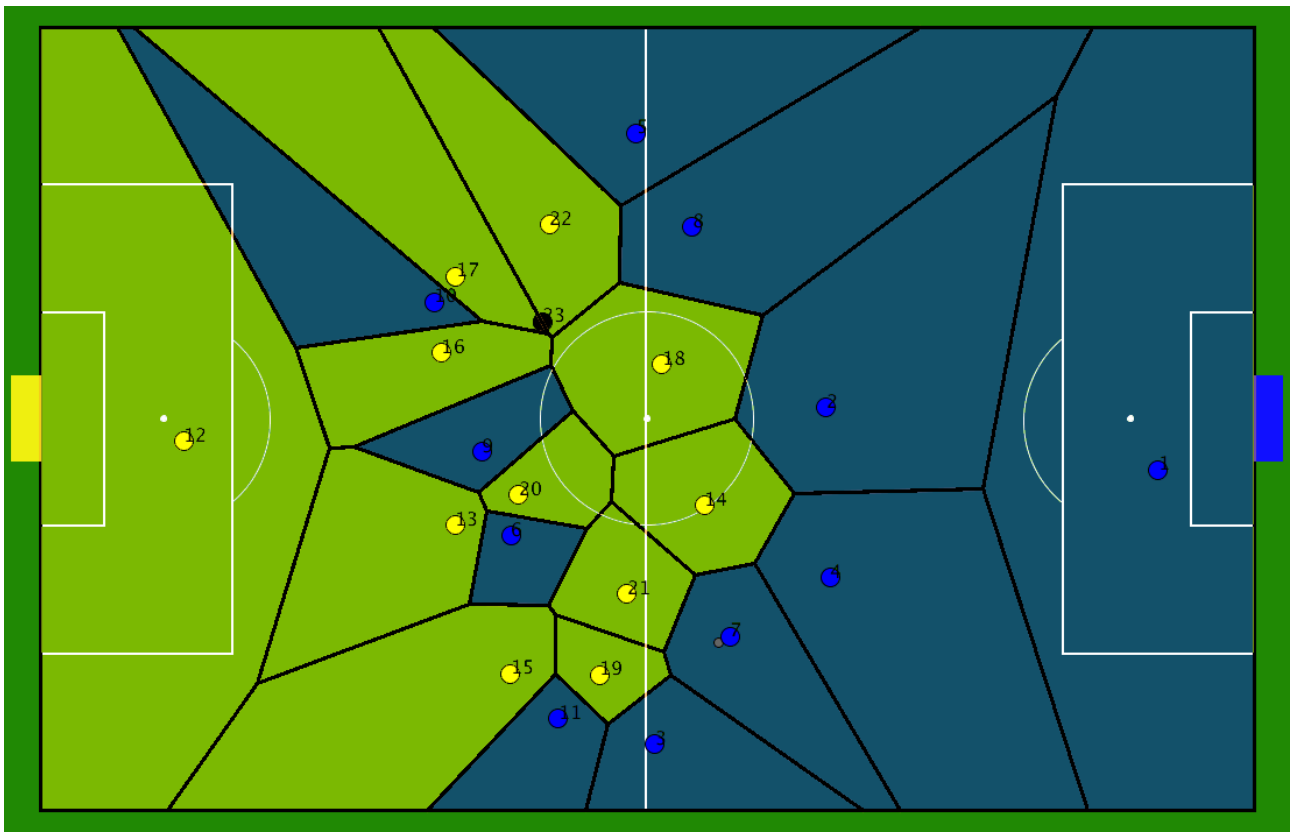

Figure 4.3: Voronoi surface teams.

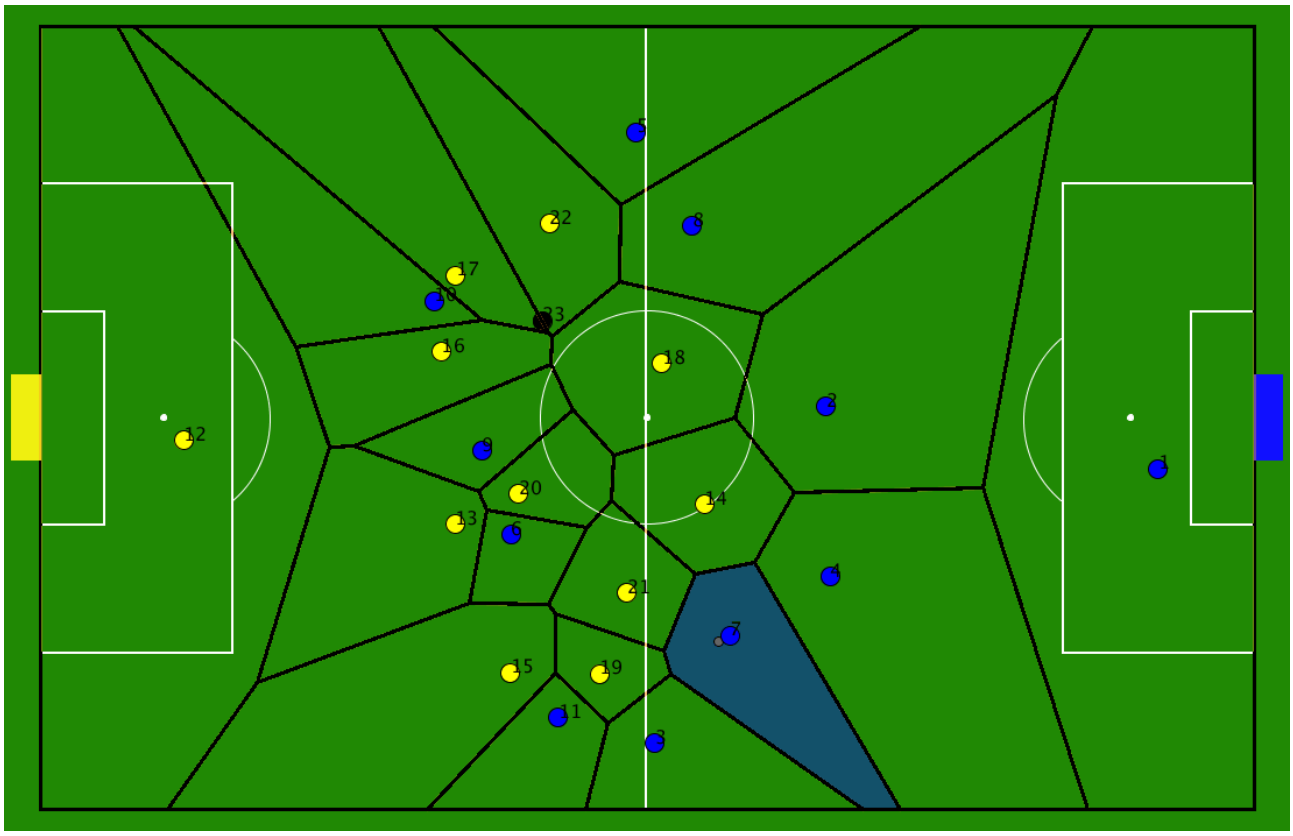

Figure 4.4: Voronoi surface player in possession and ball.

Voronoi surface of player in possession (1 feature)

The player in possession is an actor and has his own voronoi cell as represented in Figure 4.4. Because the ball is not necessarily in the voronoi cell of the player in possession another feature was made for the player in possession. The size of the voronoi cell is equal to the dominant region of the player in possession. The surface of the voronoi cell can give an indication of the pressure level on the player in possession.

Spatial pressure r meters (7 features)

Spatial pressure [HGCE15] deals with the dominance of the team in possession in a circle around the player in possession. The region in the circle dominated by the team in possession defines the pressure. If the team in possession is dominant in 100 per cent of the circle, the pressure can be seen as very low. If the team in possession is dominant in only 30 per cent of the circle the player is under high pressure. Figure 4.5 represents two of these different situations. The dominance of the circle of course depends on the size of the circle, therefore the radius r of the circle can either be 2, 3, 4, 5, 6, 7 or 8 meters. This results in seven features which define the spatial pressure for the different circle sizes.

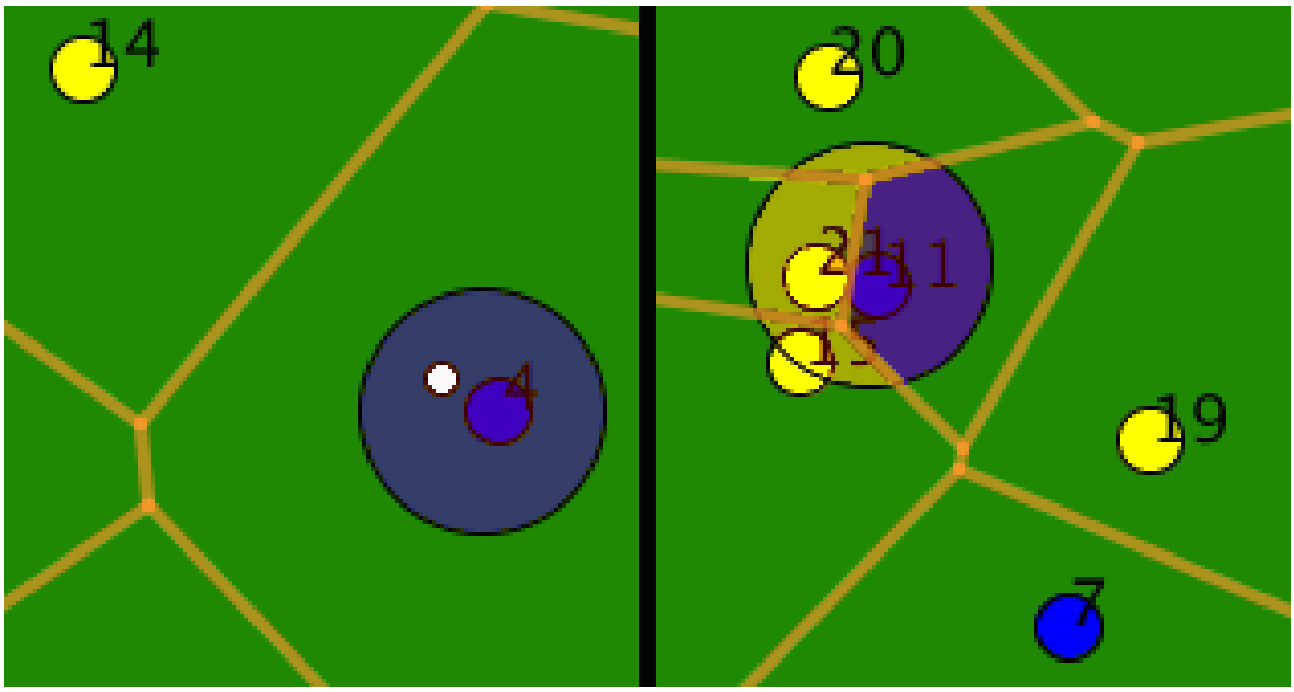

Figure 4.5: Spatial pressure, radius is 6, low pressure versus high pressure.

4.7 Possible passes

An important factor in football is the number of players able to receive the ball at their current location. If the player in possession has a lot of teammates able to receive the ball it is expected that he is less likely to loose the ball. Therefore the ability to receive the ball is tried to be captured by making ten triangles between the player in possession, the teammates and the opponent forming the smallest triangle. The basis is the length of the possible pass, and the height is the perpendicular to this line. The intersection of these lines is where the ball could possibly be intercepted. Figure 4.6 represents two examples of two triangles where the red dot could be an interception point.

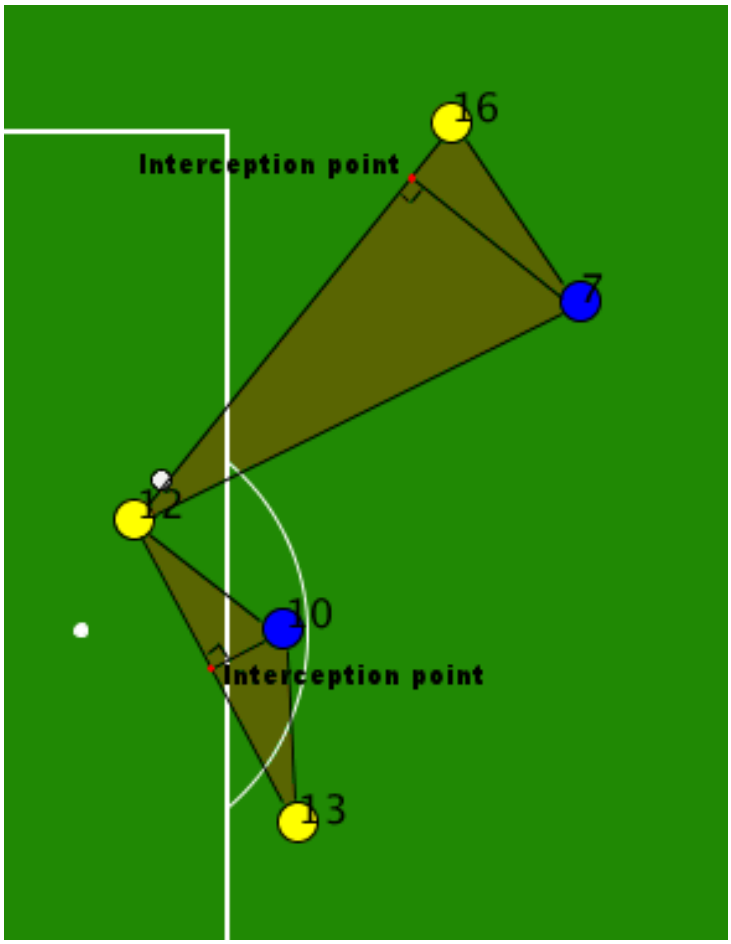

Figure 4.6: Possible pass examples.

Surface of triangle (1 feature)

One feature is the surface (red area in Figure 4.6) of the triangle normalized by the Formula 4.6 with the surfaces of the triangle as the input. The feature tries to capture the ability to receive the ball by the team mates. If the sum of the normalized surfaces is high it is more likely that a lot of players are able to receive the ball compared to when the sum of the normalized surfaces is low.

$$normalize(x) = \sum_{x=1}^{10} 1 - \frac{1}{x+1} \tag{4.6}$$

Ratio height versus basis (1 feature)

When a pass is very long and the meters to a possible interception point are low for the opponent player it is very likely that the ball is recaptured at this point. If the other way around it is not likely that a possession switch takes place. This phenomenon is tried to be captured in this feature using the Formula 4.6 with as parameter the ratio of the walking meters to the interception point versus the ball meters (height/basis of triangle).

ratio height versus ball to interception point (1 feature)

Because a ball can travel faster then a player can run the ratio between the distance the player has to run to the interception point and the distance ball has to travel to the interception point is also relevant. If the ball has to travel a long way to the possible interception point and the player only has to do a few steps to the possible interception point, the ball is likely to be intercepted. While if it is the other way around it is less likely that the ball is recaptured. The ratio is again used as a parameter for the Formula 4.6 so that the sum again can tell something about the ability to recapture the ball.

Chapter 5

Target

As described in Chapter 2, a target is needed to do subgroup discovery [FF99, KZ02]. This chapter describes three different definitions of a ball possession switch. For every definition eight different intervals are made. This method provides us with 24 targets, the different targets are represented as 0 or 1.

5.1 Definitions used

Three different definitions of a ball possession switch are used: change of ball possession calculated using the method described in the data Chapter 3, possession gain extracted from the event data and possession loss also extracted from the event data.

5.1.1 Ball possession switch

The main target depends on which team is in ball possession. This is calculated using the ball possession during the match explained in the data Chapter 3. The ball possession therefore is based on the event data and the player who touched the ball for the last time. The ball possession is recalculated for every frame which results in an array of ones and twos. By looping over the array, a change of ball possession can be found by comparing the current and the next frame. If the ball possession from the current frame differs from the next frame, a possession switch is found and the target can be set to 1 for the current frame. When the ball possession from the current frame is equal to the next frame a 0 is stored for the current frame.

5.1.2 Possession gain

The second target which might be interesting is the possession gain target. This target is provided by the event data and represents a ball possession gain from either team 1 or team 2. Just like the target ball possession switch a possession gain is represented with a 1, does the definition of the event data not contain possession gain a 0 was stored for this frame. This target is registered with a computer operated by human, the downside of this is that the target might not be precisely timed. The other confusing part of this target is that not for every occurrence of possession gain a possession loss is registered.

5.1.3 Possession loss

The third target which might be interesting is the possession loss target. The possession loss is again represented with a 0 and a 1. For this target, the same downsides apply as for the possession gain target.

5.2 Intervals of the targets

Not only the exact moment of a ball possession switch is interesting. Therefore the range and interval of the targets are expanded in eight different ways visualized in Figure 5.1 and enumerated here:

- Exact moment / target 1.
- Exact moment till 0.5 seconds before / target 5.
- Exact moment till 1 seconds before / target 10.
- Exact moment till 2 seconds before / target 20.
- Exact moment till 4 seconds before / target 40.
- 1 second before till 2 seconds before / target 10-20.
- 1 second before till 3 seconds before / target 10-30.
- 1 second before till 4 seconds before / target 10-40.

The interval which end one second before the possession switch took place is made because the moment before a possession switch could also be interesting. The target inside the intervals are represented with a 1 for every frame in the interval while the targets values of frames outside the intervals are 0. The intervals are not only made for the ball possession switches but also for the possession gain and possession loss.

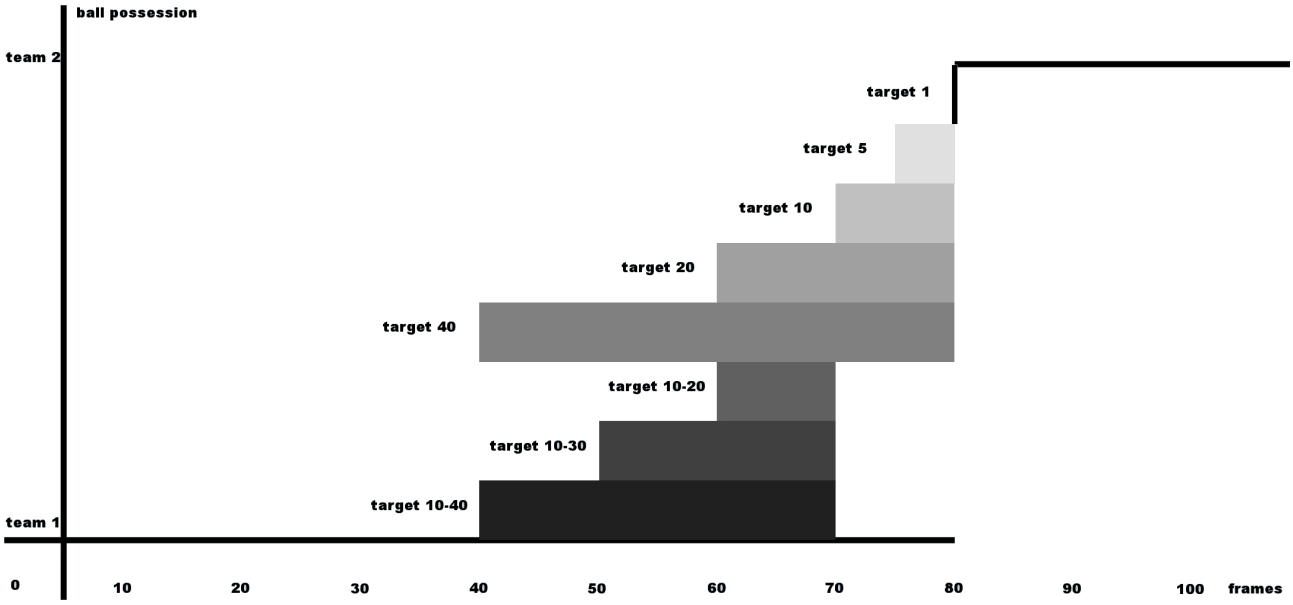

Figure 5.1: Intervals of targets.

Chapter 6

Experiments

Several experiments are done to discover which features can define a ball possession switch. In the experiments, the data set containing all the features from Chapter 4, a few selected targets from Chapter 5, and a selection of frames from Chapter 3 are used. A discussion will follow after the results are presented.

6.1 Experiments setup

In this section the experimental setup is explained. During the experiments different targets are used, only a selection of the frames are used, and the parameters for statistically significant subgroups are used in Cortana.

6.1.1 Data used

The experiments are done using different slices and dices from the complete data set of all the frames, the features and the targets. This is done because some frames and features are not reliable enough for the input. The unreliable data is due to the frames where the ball is not in sight of the camera, one or more players are not in sight of the camera, players running faster than 36 kilometers per hour or a ball going faster than 144 kilometers per hour. The maximum speed is chosen because the only one sprinting faster is probably Usain Bolt during his world record sprint at the 100 meters in 2009 [IAA16]. The top speed of a ball is set to 144 kilometers per hour because no one from the Eredivisie participants or any other dutch team made it into the top 10 of fastest shots ever [Spo15]. The the number 10 was a shot of 129.5 kilometers per hour.

The size of the full data set is 592.203 frames, all the frames are snapshots of one of the 9 matches mentioned in Chapter 3. In 226.002 of these frames, not all 22 players playing are registered by the camera. The cause of missing player values could be: it was a frame that took place during the break, it was a frame where one or more players were not registered by the camera or one or more players received a red card and were send of. In 148.181 of these frames, the ball is not registered by camera. This could be because the ball was not in sight of the camera or it was during the break. In 28.299 frames, the ball or players their speed happened to be to fast. After deleting all the unreliable frames, we are left with a data set consisting of 288.607 frames which is 48.73% out of the 592.203 frames.

After the deletion of all the unreliable frames, the data set contains 1171 possession switches defined following the definitions of target 1, 959 possession losses according to the event data and we have 968 possession gains according to the event data. The full data set contained 1671 possession switches, 1459 possession losses and 1416 possession gains.

6.1.2 Tool used

The tool we use to analyse the data set is Cortana [DFK16, MK11], Cortana is a generic data mining tool specially developed for subgroup discovery [FF99, KZ02]. This open source application can deal with several different data types as binary, ordinal, numeric and nominal values. Cortana also provides functionality for computing the threshold. This functionality makes sure that every subgroup is statistically relevant.

6.1.3 Parameters used

For every experiment with the different targets, mostly the same parameters are used to get subgroups that are significantly relevant. Four different fields should be filled: the data set explained before, the target concept, the search conditions and the search strategy. The parameters are set with the goal to reach the biggest area under the curve. The parameters are chosen after a process of trail and error. Figure 6.1 represents a screen shot of Cortana with the parameter settings set for the first experiment.

Target concept The target concept and parameters can be found in Table 6.1. The primary targets in this table are set respectively for every distinct experiment.

The parameters used for the target concept are all binary as described in Chapter 5 therefore the target type is single nominal. The target value we are looking for is always 1 because we want to analyse possession

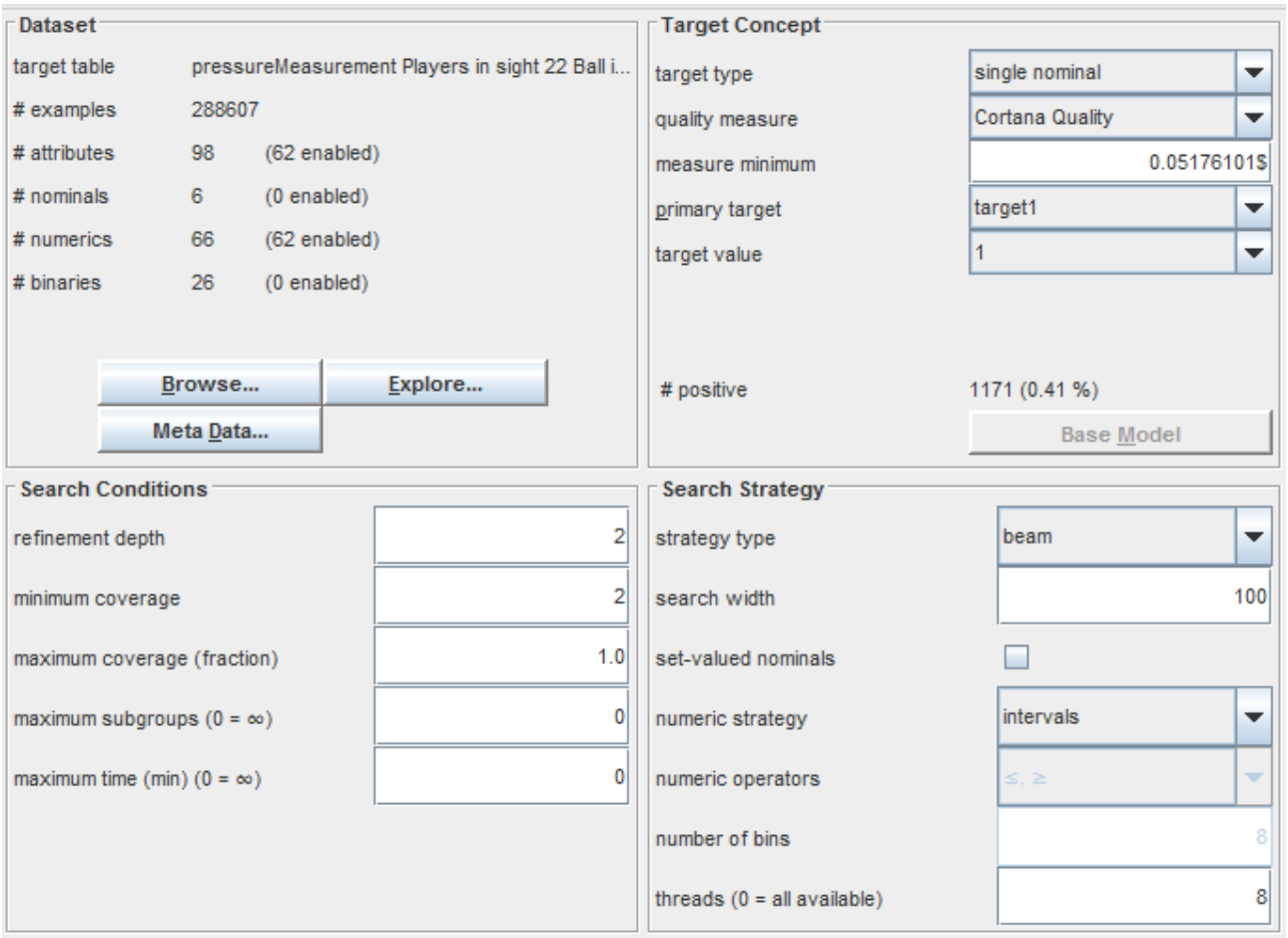

Figure 6.1: Cortana parameters experiment 1.

Search parameter	Value
Target type	single nominal
Quality measure	Cortana quality
Measure minimum	0.1
Primary target	target 1, target 10 gain, target 10-20 loss
Target value	1

Table 6.1: Target concept.

switches. The quality measure is Cortana Quality and the measure minimum is set to 0.1 so only the statistical relevant subgroups will be found. The measure minimum can be lowered with the compute threshold button but the subgroups so close to the baseline are not the ones we are looking for any way.

Search strategy The parameters from Table 6.2 are set for the search strategy parameters for all the experiments.

Search parameter	Value
Strategy type	beam
Search width	100
Numeric strategy	interval

Table 6.2: Search strategy.

The numeric strategy are parameters to pay attention to. The numeric strategy is changed from the default best-bins strategy to the interval strategy. This is done because this interval strategy comes up with intervals instead of only $\leq$ or $\geq$ than a particular size. Because the interval numeric strategy is chosen the number of bins does not matter any more.

Search conditions The search condition parameters are set to the values in Table 6.3

Search parameter	Value
Refinement depth	2
Minimum coverage	2
Maximum coverage (fraction)	1
Maximum subgroups	∞
Maximum time (min)	∞

Table 6.3: Search conditions.

The refinement depth is set to 2 so the subgroup discovery and combines two or less features that indicate a ball possession switch. This can be easily extended to 3 but this does not significantly improve the quality of the subgroups.

6.2 Results

Three different experiments are done, the main difference between the experiments are the targets. For every target the most important features are summed, the most important subgroups with their quality are summed and the ROC curve belonging to the subgroups is given.

6.2.1 Target 1

With a refinement dept of 1, the five subgroups with the highest quality are presented in Table 6.4. Like described in the experimental setup, the interval setup is used. This way interesting intervals of important features can be found.

quality	positives	feature	interval
0.592	1,063	ballDistTo1ClosestPlayersDefTeam	(0.191, 4.362]
0.571	1,112	defPlayersWithin5Meters	(0.0, ∞)
0.439	958	ballDistTo2ClosestPlayersDefTeam	(1.483, 14.920]
0.399	865	ballDistTo3ClosestPlayersDefTeam	(4.959, 26.814]
0.375	877	ballDistTo4ClosestPlayersDefTeam	(9.18683, 45.1135]

Table 6.4: Important variables target 1.

The feature *ball distance to the* 1 *closest players of the defending team* has the highest quality. This is no big surprise because of the choices made according to the definition of ball possession in Chapter 3. When one or more players are closer to the ball than five meters it is also very likely that a possession switch takes place. The feature *ball distance to the* 2 *closest players of the defending team* has the interval ~ 2 to ~ 15. The feature *ball distance to the* 3 *closest players of the defending team* has the interval ~ 5 to ~ 27. The feature *ball distance to the* 4 *closest players of the defending team* has the interval ~ 9 to ~ 45. These intervals can be interesting in football perspective because they can be used in football practice. The application of this is discussed later.

ROC curve The ROC curve is calculated with a refinement depth of 2. The subgroups forming the ROC curve are presented in Table 6.5 and the ROC curve can be found in Figure 6.2. The Area Under the Curve (AUC) is 0.846. This means that the qualities of the subgroups on the ROC curve are quite good.

FPR	TPR	Condition 1	Condition 2
0.095	0.513	distPlayerInPossessionAndBall in (2.851, 73.614]	ballDistTo1ClosestPlayersDefTeam in (2.017, 4.479]
0.313	0.906	ballDistTo1ClosestPlayersDefTeam in (0.191, 4.363]	ballDistTo3ClosestPlayersAttTeam in (7.347, ∞)
0.314	0.908	ballDistTo1ClosestPlayersDefTeam in (0.191, 4.363]	ballDistToOpponentGoal in (1.242, 107.954]
0.377	0.950	defPlayersWithin5Meters in (0.0, ∞)	ballDistToOpponentGoal in (1.242, 107.954]

Table 6.5: Subgroups on ROC curve target 1 (AUC 0.846).

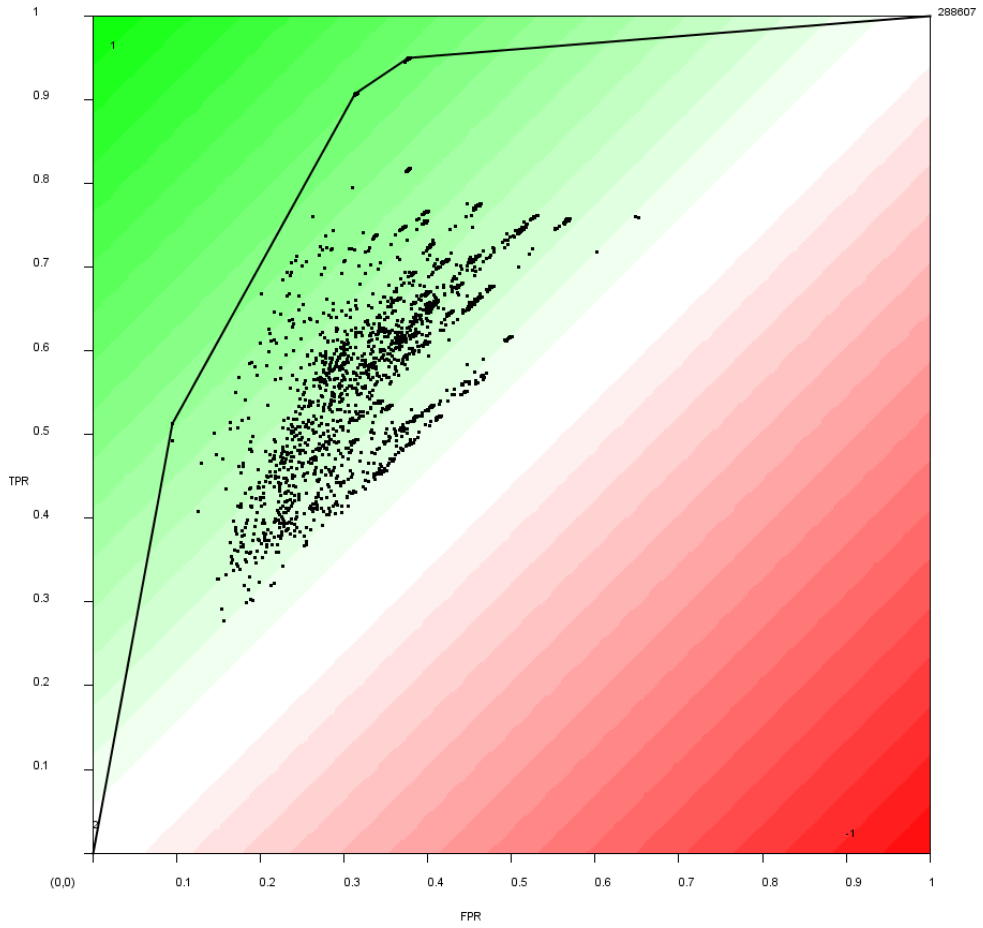

Figure 6.2: ROC curve target 1 (AUC 0.846).

6.2.2 Target10Gain

With a refinement dept of 1, the three subgroups with the highest quality are presented in Table 6.6. Like described in the experimental setup, the interval method is used. This way interesting intervals of the important features for this target can be found.

quality	positives	feature	interval
0.254	6,247	voronoiSurfaceAttTeam	(1,021.89, 3,857.76]
0.254	6,247	voronoiSurfaceDefTeam	(3,282.22, 6,118.07]
0.220	4,727	ballDistToAttTeam	(100.384, 251.936]

Table 6.6: Important variables target 10 gain.

The feature *voronoi surface of the attacking team* has the highest quality. The interval of this feature is 1022 ($\sim$ 15% of the field) to 3858 ($\sim$ 57% of the field). This means that the dominant region of the attacking team at time of a possession gain is in most cases between 15% and 57% of the field. This interval is not a logic interval, because it usually is the other way around. Right after the moment of a possession gain for the defending team, the previous defending team becomes the attacking team and the other way around. The attacking team (the previous defending team) is usually still dominant on their own half while the other half

will be dominated by both the teams with a separation of about $50 - 50$. This way in most cases the interval of the feature *voronoi surface of the attacking team* can not be 15% to 57% for a possession gain.

ROC curve The ROC curve is calculated with a refinement depth of 2. The subgroups forming the ROC curve are presented in Table 6.7 and the ROC curve can be found in Figure 6.3. The AUC is only 0.661, which means that the qualities of the subgroups on the ROC curve are not so high.

FPR	TPR	Condition 1	Condition 2
0.124	0.318	ballDistTo3ClosestPlayersAttTeam in (3.410, 22.220]	teamOpponentDistRatio6competitors in (0.410, 1.212]
0.147	0.361	ballDistTo4ClosestPlayersAttTeam in (9.719, 39.730]	teamOpponentDistRatio6competitors in (0.410, 1.210]
0.155	0.377	ballDistTo5ClosestPlayersAttTeam in (14.628, 59.998]	teamOpponentDistRatio6competitors in (0.410, 1.210]
0.193	0.446	ballDistTo6ClosestPlayersAttTeam in (20.695, 81.995]	voronoiSurfaceAttTeam in (1,021.55, 4,563.82]
0.279	0.563	voronoiSurfaceAttTeam in (1,021.89, 3,857.76]	playerDistTo6ClosestOpponents in (13.920, 59.138]
0.311	0.590	voronoiSurfaceAttTeam in (1,021.89, 3,857.76]	possessionTeamOpponentTeamRatio15m in (0.4, 4.0]
0.381	0.644	voronoiSurfaceAttTeam in (1,021.89, 3,857.76]	ballDistToOpponentGoal in (6.671, 104.91]
0.384	0.646	voronoiSurfaceAttTeam in (1,021.89, 3,857.76]	directOpponentDist in $(-\infty, 142.427]$

Table 6.7: Subgroups on ROC curve target 10 gain (AUC 0.661).

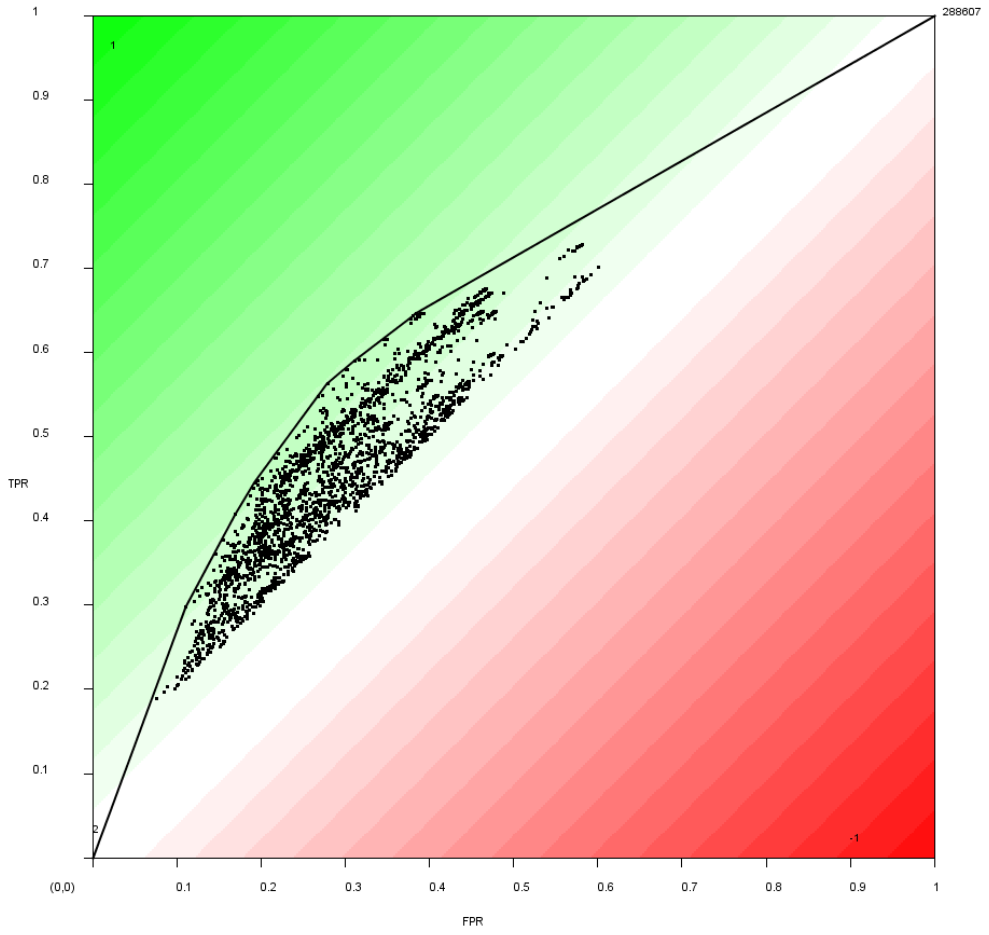

Figure 6.3: ROC curve target 10 gain(AUC 0.661).

6.2.3 Target 10-20 loss

With a refinement dept of 1, the three most important features are presented in Table 6.8. Like described in the experimental setup, the interval method is used. This way interesting intervals of the important features for this target can be found.

quality	positives	feature	interval
0.268	6,217	ballDistTo3ClosestPlayersDefTeam	(2.744, 28.490]
0.266	6,506	ballDistTo4ClosestPlayersDefTeam	$(-\infty, 47.423]$
0.264	7,293	ballDistTo2ClosestPlayersDefTeam	(1.189, 17.960]

Table 6.8: Important variables target 10-20 loss.

The feature *ball distance to the 3 closest players of the defending team* has the highest quality. This feature states that in the second before a possession loss the distance to the 3 closest players from the team trying to recapture the ball lies in the interval of $\sim$ 3 and $\sim$ 28. The feature *ball distance to the 4 closest players of the defending team* has an interval of $-\infty$ to $\sim$ 47. The minus infinity can be seen as 0 because negative values for distances do not exist on a football field. The feature *ball distance to the 2 closest players of the defending team* has an interval of $\sim$ 1 to $\sim$ 18. These intervals are interesting because we have seen comparable intervals before for target 1. This again can be used in football practice and will be discussed later. The downside of the intervals for target 10-20 loss is that the quality of the subgroups found is not so high.

ROC curve The ROC curve is calculated with a refinement depth of 2. The subgroups forming the ROC curve are presented in Table 6.9 and the ROC curve can be found in Figure 6.4. The AUC is 0.662, which means that the qualities of the subgroups on the ROC curve are again not so high.

FPR	TPR	Condition 1	Condition 2
0.082	0.184	voronoiSurfaceDefTeam in (3,945.96, 6378.08]	defPlayersWithin5Meters in (0.0, 4.0]
0.112	0.242	voronoiSurfaceDefTeam in (3,945.96, 6378.08]	ballDistTo4ClosestPlayersDefTeam in (16.242, 55.707]
0.214	0.415	freePlayers7meters in $(-\infty, 3.0]$	defPlayersWithin15Meters in (2.0, 8.0]
0.350	0.619	ballDistTo3ClosestPlayersDefTeam in (2.744, 28.490]	freePlayers3meters in (1.0, 9.0]
0.366	0.639	ballDistTo3ClosestPlayersDefTeam in (2.744, 28.490]	playerDistTo6ClosestOpponents in (16.558, 91.123]
0.471	0.745	ballDistTo2ClosestPlayersDefTeam in (1.189, 17.960]	ballDistTo2ClosestPlayersAttTeam in (2.087, 25.301]
0.474	0.748	ballDistTo2ClosestPlayersDefTeam in (1.189, 17.960]	ballDistTo1ClosestPlayersAttTeam in (0.015, 10.421]

Table 6.9: Subgroups on ROC curve target 10-20 loss (AUC 0.662).

6.3 Results from a football perspective

From a football perspective, it is very obvious that the closest defending player should be close to the ball to recapture ball possession. But it turned out that other players close to the ball is also an important aspect. For example if we take a look at the exact moment a possessions witch occurs, the distance between the 2 closest players from the defending team and the ball should be in the range of $\sim$ 1.5 meters $\sim$ 15 meters.

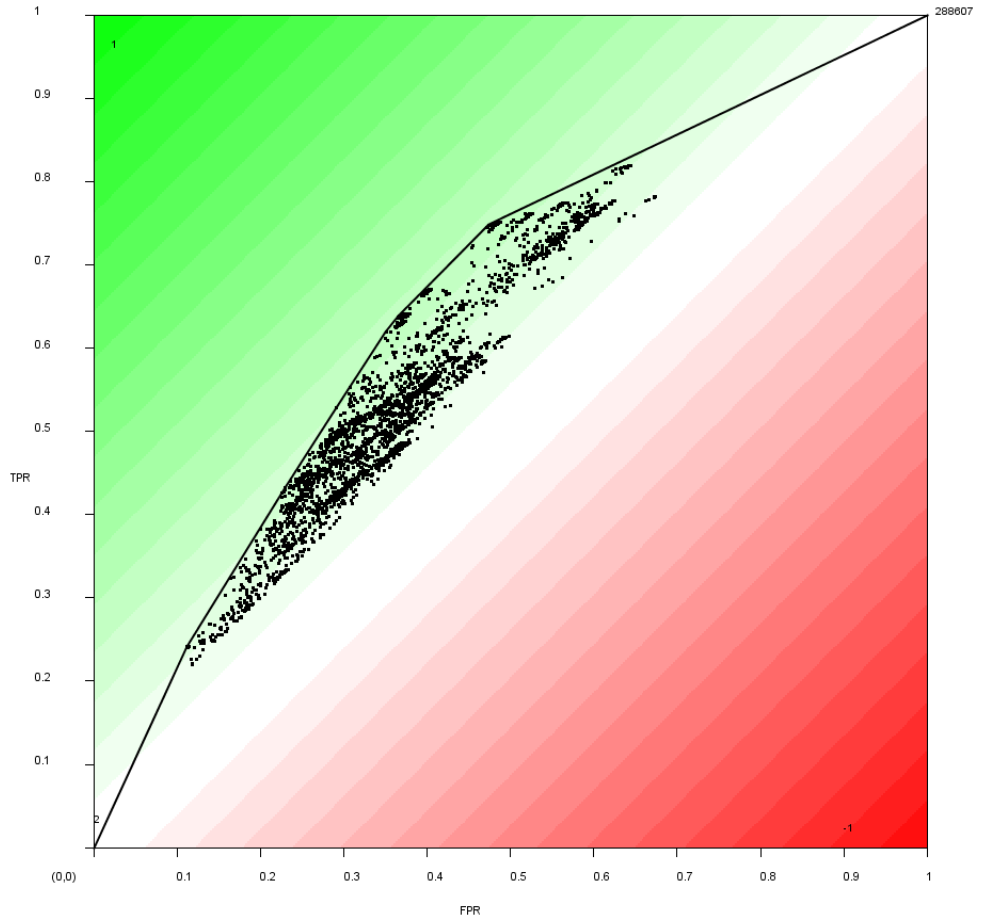

Figure 6.4: ROC curve target 10-20 loss (AUC 0.661).

The distance between the 3 closest players from the defending team and the ball should be in the range of $\sim$ 5 meters to $\sim$ 27 meters and in 877 out of the 1171 cases, the distance between the ball and the closest 4 players from the defending team were in a range of $\sim$ 9 meters to $\sim$ 45 meters for a possession switch to occur.

Comparable subgroups were found a second before a possession loss took place. The subgroups found for this target turned out to be very similar to the exact moment of a possession switch. Therefore we can conclude that the subgroups found are relevant for possession losses and possession switches. The quality of these subgroups were not as high as the quality of the subgroups found for the exact moment. That is why the subgroups with their intervals for target 1 are more trustworthy.

The intervals of target 1 could for example be used in football practice. Make a little field with the size 7 by 7 so the distance between the ball and the closest three defenders can only in the worst case scenario be bigger than 27 meters. Let the players play a little match of 3 against 3. This could also be done with different field sizes for matches of 2 against 2 and 4 against 4.

6.4 Discussion

The subgroups found are probably not 100% reliable because of the definition of a possession switch, the quality of the data and the quality of the features. Which of these three causes plays the biggest role is unknown. Why these factors play a role is shortly explained in the following sections.

6.4.1 Definition of possession switch

The team in possession explained in Chapter 3 is the guidance for most features. Because of the team possession, the features belonging to the frame where a possession gain and possession loss (both extracted from the event data) took place are wrong in most cases. The features do not take a false timing in consideration so the features in these frames are wrong. The team in possession could have changed already in one of the frames earlier or later. The features in the frame corresponding to a possession loss or possession gain looks like noise in the data.

Target 1 is also not completely right because this target does take fouls or possession switches due to a ball outside the lines in consideration.

6.4.2 Quality of the data used

The data already has a lot of gaps, missing values, speed issues, and timing issues. After filtering these moments out, 48% remains but it is not said that this remaining data is 100% reliable. This is due to a missing *z-coordinate* for the ball, and the confusion between players that became visible while visualizing the match.

6.4.3 Quality of the features

During this research, a lot of feature engineering is done but of course not all the possible features were calculated. There might be features that can predict possession switches much better that were not calculated in this research. It could also be that the possession switches are not 100% predictable because every match is different and because of this, football is very unpredictable.

Chapter 7

Conclusions

Before we can answer the research question, and tell which features are most relevant for possession switches, the sub-questions should be answered. The data available for our research were nine Eredivisie matches from the first half of the season 2015-2016. The data from these matches were visualized and turned out to be unreliable and incomplete in more than half of all the frames. 62 features were extracted from the data that can give an indication of pressure and hopefully would catch possession switches. The results from the subgroup discovery where not very eye opening, but several things became clear. The distance between the four defending players and the ball turned out to be the most important variables. The subgroups found but did not cover all possession switches and contained a lot of false alarms. This means that the difference between some possession switches and the possession switch in the subgroups were to big. The false alarms occurred because of situations which looked like a possession switch without a possessions switch actually happening.

7.1 Relevant features

To answer the research question stated in the introduction of the research: *Using an algorithmic approach that analyses possession switches during football matches, which variables have the highest influence on these moments?* The most relevant variables are the ones that define the distance between the ball and the closest defending players. This does not come as a surprise but it is information we can learn from. The total distance between the ball and the closest 1, 2, 3 and 4 defending players should respectively fall in the interval of ~ 0 meters to ~ 4 meters, ~ 1.5 meters to ~ 15 meters, ~ 5 meters to ~ 27 meters and ~ 9 meters to ~ 45. This can be used in football practice so the players can train under these conditions.

7.2 What can be done after this research

This research could be done again after more accurate data is available. When more accurate data is available we will know which features truly are responsible for the possession switches. Perhaps new features can be found which might be able to predict possession switches more accurately. A second thing which could be done after this research is making a model to define pressure. This can probably be done with the variables responsible for possession switches. This way training practices could be made to practice under a self chosen pressure level.

Bibliography

[Cal16] Michael Caley. Premier league projections and new expected goals, 2016.

[Dav16] Martin Davis. Class VoronoiDiagramBuilder, 2016.

[DFK16] Wouter Duivesteijn, Ad J. Feelders, and Arno Knobbe. Exceptional Model Mining. *Data Mining and Knowledge Discovery*, 30(1), 2016.

[FF99] Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2), 1999.

[GE06] Isabelle Guyon and André Elisseeff. *An Introduction to Feature Extraction*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[GH16] Joachim Gudmundsson and Michael Horton. Spatio-temporal analysis of team sports - A survey. *CoRR*, abs/1602.06994, 2016.

[HCK16] Miao He, Ricardo Cachucho, and Arno Knobbe. Football players performance and market value. 2016.

[HGCE15] Michael Horton, Joachim Gudmundsson, Sanjay Chawla, and Joël Estephan. *Automated Classification of Passing in Football*. Springer International Publishing, Cham, 2015.

[IAA16] IAAF. Usain Bolt, athlete profile, 2016.

[Kuh55] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2), 1955.

[KZ02] Willi Klösgen and Jan M. Zytkow, editors. *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, Inc., New York, NY, USA, 2002.

[MG02] S. J. Mason and N. E. Graham. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q.J.R. Meteorol. Soc.*, 128(584), July 2002.

[MK11] Marvin Meeng and Arno Knobbe. Flexible enrichment with Cortana – Software Demo. *Proceedings of BeneLearn*, 2011.

[Opt16] Opta. Blowing up everything you know about sports, 2016.

[PS85] Franco P. Preparata and Michael I. Shamos. *Computational Geometry: An Introduction*. Springer-Verlag New York, Inc., New York, NY, USA, 1985.

[Spo15] Silly Season Sport. Top 10: Hardest shots ever recorded in football, 2015.

[UC08] Graham Upton and Ian Cook. *A Dictionary of Statistics*. Oxford University Press Print, 2008.

[WF05] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[Wor16] Tom Worville. What's the best football analytics piece you've ever read?, 2016.