Universiteit Leiden
The Netherlands

Opleiding Informatica

Characteristics of dangerous passes in soccer

at Women's EURO 2017

Jeroen Rook

Supervisors:
Dr. Arie-Willem de Leeuw & Dr. Arno Knobbe

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2018

Abstract

Within sport analytics, performance indicators are important measures for comparing performance between teams and to investigate tactics. In soccer, the amount of passes and the pass accuracy are examples of such indicators. These, however, tell nothing about the quality of the passes. More specific indicators are required to represent the quality of the passes. Before such indicators can be defined, we first need to define different types of passes and know their characteristics. These characteristics could also be useful for making strategies. Out of 6 matches, which have been captured at the Womens EURO 2017, we extracted all passes. As a first exploration, we look at the characteristics of dangerous passes by performing manual feature engineering. Subgroup discovery and rule based target values are used in order to identify characteristics of dangerous passes. Several subgroups holding a significantly higher percentage of dangerous passes were discovered. Out of these subgroups, we define characteristics. For example, passes which are made on the other half and are played backwards more than $5.7m$ are more dangerous.

Acknowledgements

I would like express my profound gratitude to my thesis supervisor dr. Arie-Willem de Leeuw for his guidance and comments throughout all the stages of this project and the detailed feedback in the writing of this thesis.

I would also like to thank my second supervisor dr. Arno Knobbe for giving comments on my thesis progress as well for giving me a detailed explanation into the methods and tools needed for this thesis.

Finally, I want to express my gratitude to all researches, fellow students, friends and family, who provided sharp remarks and valuable comments during this project.

Contents

Chapter 1

Introduction

In sports the main goal is to perform better than your opponents. Therefore, it is important for athletes to optimise all aspects that have an impact on the final performance. Optimising the physiological capacities is probably the most important aspect that is present in almost all sports. For example, the cyclists who attended the first Tour the France tried to optimise their performance by cycling long distances each day. Over the years new knowledge resulted in other training schemes, which cover less distance and give better performances.

A physiological advantage is only one part of the equation of winning in sports. Technical and tactical capacities are of equal importance for the overall success of an athlete or a team. Knowing how to quantify these capacities is a challenging and complex task. One way to do this is by using notational analysis. Generally, several statistics are notated on which a comparison can be made between different athletes or teams, but it can also give better insights on which strategies are more valuable. For example, at the 1990 Fifa World Cup, successful teams performed better in converting possession into "shots on goal" than unsuccessful teams [1]. Here, a team was classified *successful* if it made it into the quarter final, and *unsuccessful* if they were among the first round losers. In this case, the variables only gave a global picture of the match statistics and did not describe it in much detail. Nowadays, video tracking makes it possible to notate spatiotemporal variables for objects and athletes. With this data, more accurate measurements can be made for the performance of a team or athlete.

In soccer, the performance is based on multiple aspects, such as passes, tackles, goals, field position and time in possession [2]. In order to measure the technical performance, for example, the percentage of passes which reach a teammate and the number of shots on target is important. For tactical performance, some examples are passes/possession and passing distribution. These are specific examples of performance indicators. All of these examples are based on passes. However, they are used as an asset rather than as a measure itself. In this bachelor thesis, we will try to get more insights on how contributory individual passes are.

1.1 Research question

The following question will be answered in this thesis:
What are characteristics of dangerous passes in soccer at the WEURO 2017?

1.2 Related work

Passes have been an important factor for measuring performance. However, with notational analysis it is mainly used in a certain ratio with other factors, such as corners or shots, or to a normalised variable [2]. Hughes et al. [1] for example showed that significantly more shots were made when the passing sequences were longer.

There are some articles which try to value a pass based on the spatiotemporal variables. Horten et al. used spatiotemporal data in order to classify passes [3]. The labelling of these passes was carried out by experts. Fairly good results were achieved, but they did not gain much insight on the characteristics of a good pass. More recently, Rein et al. looked into the characteristics of a pass using spatiotemporal data [4]. Passes which occurred within a particular third on the pitch or that went from one third to another were compared. The sections were defined as *back,middle* and *front*. They concluded that passes which changed the space control in the attacking third as well as the number of outplayed defenders are valuable measures for the success of a pass.

1.3 Overview

This thesis is organised as follows. In chapter 2, we define the domain accompanied with the terminology involved. Also the tools are explained here. In chapter 3 we look at the quality of the dataset and describe the pre-processing steps which resulted in a set of individual passes. Chapter 4 explains all the different types of features we constructed. With all these features for each pass we have a set on which we can perform subgroup discovery in order to find characters tics. These experiments and results are described in chapter 5. In the last chapter, we conclude on these results and give possibilities for future research.

Chapter 2

Preliminaries

2.1 Soccer

Soccer, also known as association football, is a worldwide known sport and is practised all over the world. In general, all official soccer matches live by the Law of the Game [5] published by International Football Association Board (IFAB). In general, two teams, each consisting of 11 players, play on a pitch and try to put the ball in the opponent his goal. Without using their arms, the players can dribble with the ball or pass it to another player. The team with ball possession is called the attacking team and the other team is the defending team. Figure 2.1 shows the pitch with the terminology used for the lines and zones.

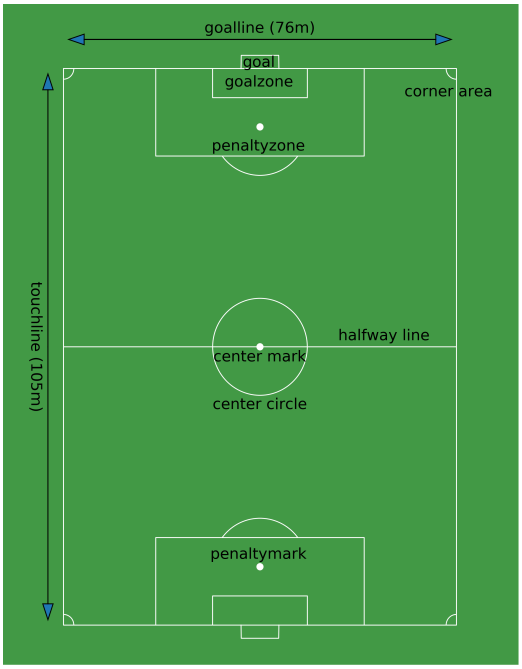

Figure 2.1: Soccer field with markings and zones

2.2 Subgroup discovery

Subgroup discovery is the discovery of subsets of the data which have a difference in distribution of a single target attribute compared the the whole dataset [6,7]. Where other classification algorithms make models to fit the whole dataset, subgroup discovery does not. The task of the subgroup discovery algorithm is to find all subgroups, using a certain search strategy, and determining the quality using a quality measure. In order to be considered a subgroup they must fit specified inductive constraints such as minimum coverage and minimum quality.

2.2.1 Search strategy

There are various strategies for finding subgroups. For each feature a condition can be set on which a subgroup can be formed. With nominal features the condition can be, for instance, *Passdirection = "Forward"*. All passes holding this condition, are part of the subgroup. The amount of conditions is the number of different values of that feature. For numerical features the possible conditions are more diverse. A condition can be $\leq$, $\geq$ or $=$ to a certain number within the range of the feature. These numbers can be determined by binning the feature range or by finding the best number within the range. The size of possible conditions is based on the binning width and types of conditions.

When a subgroup is formed out of 1 condition it has a depth of 1. When more conditions are added, the depth increases linearly. The possible configurations increase exponentially when the depth is increased. A depth of 3 for 77 numerical features would give a huge search space. An exhaustive search is not recommended since it requires a lot of computational power. Therefore search strategies such as a beam search, best-first search and depth-first search can be used to refine more promising solutions first and in this way leave out solutions which lead to a bad solution.

2.2.2 Quality measures

Once a subgroup is found, it can be compared to the global target value. For nominal targets, a confusion matrix can be used to represent the coverage of the subgroup over the target value as is shown in Table 2.1. Using this matrix, the quality measure can be computed.

		Target	
		T	F
Subgroup	T	0.42	0.13
	F	0.12	0.33

Table 2.1: Example of a confusion matrix of a subgroup with in each cell the faction of the total items.

The value of this measure determines the quality of the subgroup. One of these quality measures is weighted relative accuracy (WRAcc) [8] and is used as a quality measure for binary targets. The weight represents

the coverage of the subgroup (G) over the total population (S). The relative accuracy is the relative gain of positives over the total population. With binary targets this is the difference between the fraction of each target value. This can be expressed by the following equation with as input the subgroup and the target value :

$$WRAcc(G,t) = \frac{|G|}{|S|}(frac(G,t) - frac(S,t)) \tag{2.1}$$

With $frac$ defined as:

$$frac(G,t) = \frac{|\{x|x \in G \text{ and } x = t\}|}{|G|} \tag{2.2}$$

To display the performance of subgroups a ROC curve [9] can be used, of which Figure 2.2) is an example. The curve is plotted in a 2 dimensional space. The X-axis represents the false positive rate (fpr) or $1-$*specificity*. The Y-axis represents the true positive rate (tpr) or the *sensitivity*. Each subgroup has these two properties derived from the confusion matrix, displayed in Table 2.1 the specificity and sensitivity are 0.13 and 0.42 respectively. It is favourable to have a high sensitivity and a high specificity. It is not favourable to be on the diagonal of the ROC space, because that means that there is no information gain. The outer points in the space can be represented as a curve. The area below is called *Area Under the Curve* (AUC) and is the equivalent of the probability that a randomly chosen positive instance will be rated higher than a negative instance. The AUC can have a value between 0.5 and 1.0, where 0.5 is considered worthless and 1.0 is excellent. Inbetween the classifications are bad, fair and good. In Figure 2.2 the AUC is equal to 0.772. This is considered acceptable.

In order to measure the quality of numerical targets, other quality measures are needed. Continuous weighted relative accuracy (cWRAcc) is a variant on WRAcc and can be applied to numerical targets.

$$cWRAcc(G) = \frac{|G|}{|S|}(\mu^G - \mu^S) \tag{2.3}$$

2.2.3 Cortana

For our experiments we use the subgroup discovery tool Cortana [10], which is an open source software package. Cortana has multiple search strategies and quality measures implemented which allow for applying different configurations to the dataset.

2.3 Feature construction

Before we can perform subgroup discovery on captured matches, the data files needs to be transformed to a workable dataset first. From these files we extract all the passes, from which we construct a dense feature set. This is called feature engineering. Figure 2.3 shows the workflow from the raw input files to a workable

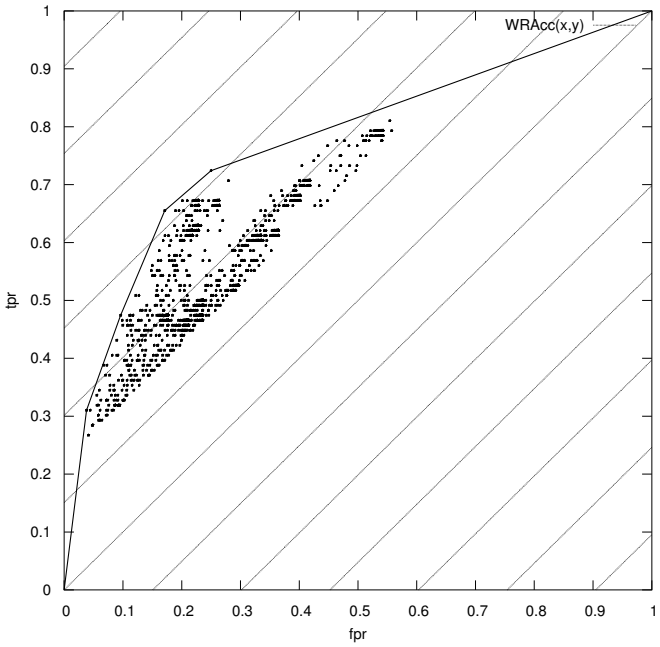

Figure 2.2: ROC curve, with on the X-axis the False Positive Rate (fpr) and on the Y-axis the False Positive Rate (fpr). The outer subgroups define the curve and the area beneath it (AUC) is a valuable measure for the performance of the subgroups as a whole.

dataset on which the experiments can be performed. Each action saves its result in its own data structure. In this way unnecessary computing is prevented. The pipeline from match to dataset is fully automated, so new matches can be added fairly easily.

All programming is performed using Python. Python has a lot of libraries which make it easy to work with. Some mentionable libraries are: *NumPy*, *SciPy*, *Pandas* and *Matplotlib*.

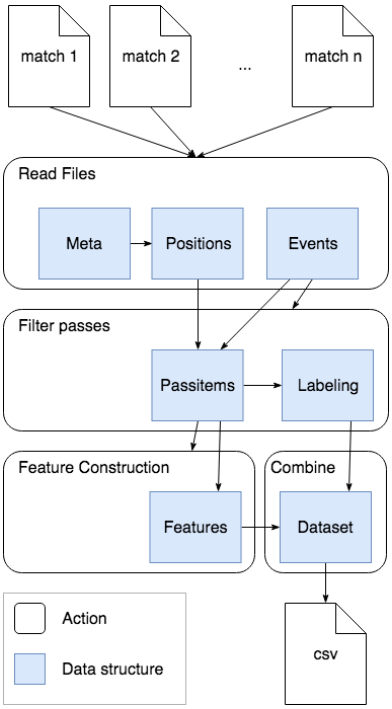

Figure 2.3: Workflow of retrieving a workable dataset. Each action outputs a new file, which act as input for another action. Inside each action the types of data which are used are in blue. The results in between each action are stored in specific data structures.

Chapter 3

Dataset

3.1 Origin

The dataset consists of 6 matches which have been played during the UEFA European Womens Championship 2017 tournament [11]. As shown in Table 3.1, the data collection consists of all matches that involved the Dutch soccer team. The matches took place in stadiums in the Netherlands over the course of three weeks.

Match	**Date**	**Outcome**
Netherlands - Norway	16 July 2017	1-0
Netherlands - Denmark	20 July 2017	1-0
Belgium - Netherlands	24 July 2017	1-2
Netherlands - Sweden	29 July 2017	2-0
Netherlands - England	3 August 2018	3-0
Netherlands - Denmark	6 Augustus 2018	4-2

Table 3.1: Recorded matches

Country	**Goals scored**	**Shots on target**	**Shots off target**
Netherlands	13	37	41
Denmark	2	10	13
Belgium	1	6	3
Norway	0	4	5
Sweden	0	3	7
England	0	7	14
Total	16	67	83

Table 3.2: Overall goals and shots per country

The last match in Table 3.1 accounts for more than one third of the total goals scored. Table 3.2 shows that the Dutch team scored 13 of all goals, which is more than 80%. For the shots this number is not that extreme and the Dutch cover more than half of them. This means that the numbers are not normally distributed and this is important to take into account when making conclusions in chapter 6.

3.2 Collection method

The raw data for each match is stored in a xml-file and consists of three different parts, meta, positional and event data.

3.2.1 Meta data

The meta data of the match describes the data which remains static during the match. It holds information on the location, such as date, stadium-name, field dimensions and information about all players. For each player, the shirt number and shoe colour are notated, but also important information such as the team to which the player belongs and a tracking number. Without these last two variables, the positional and event data would not have had much value. The meta data is purely used to compliment the other two parts where all the actual match data is stored.

3.2.2 Positional data

The positional data is captured using kinematical motion analysis [12] with three fixed cameras in the stadium. This is notated with a frequency of $10Hz$ and a step size of 0.1 m. For all moving objects, players, referees and the ball, the positions are stored as two coordinates, X and Y, for each time interval. The X-axis is parallel with the touchlines and the Y-axis is parallel with the goallines. Although a ball can travel vertically over the Z-axis, this is not tracked. The zero-point of the coordinates is on the center mark of the pitch. Based on the dimensions of the pitch, the coordinates have a position which is an element of $\{x, y \in \mathbb{N} | -525 \geq x \leq 525 \& -380 \geq y \leq 380\}$. For all 6 matches there are 8 840 086 positions in total.

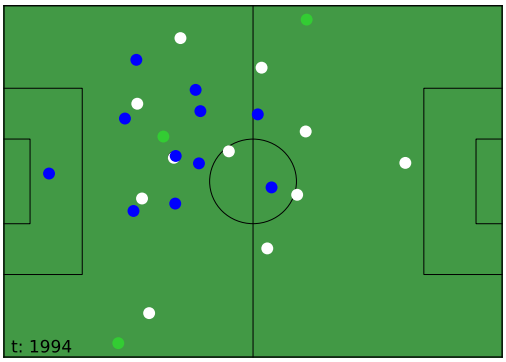

Figure 3.1: All players, referees and the ball plotted at time interval 1994. Players are shown in white and blue, black represents the ball and green dot is the referee.

With these positions for each time interval, a birds view of the situation on the pitch can be constructed. This gives an interpretable view of the situation. In Figure 3.1 team blue is possibly attacking, since all players except the goalkeeper are on the half of team white. The ball, however, is closer to a white player. The ball

could be travelling a pass trajectory or can be just intercepted by the white player. There are a lot of different interpretations for this snapshot, so additional information is required in order to clarify this. By looking at the previous frames more knowledge can be gained on what is happening. For example by using velocity vectors can indicate in which direction the teams move. Event data is also a good source for knowing what is happening on the field.

3.2.3 Event data

The event data is a list of all events in sequential order in which they have occurred. A snapshot is displayed in Table 3.3. These events can refer to something that happened to the ball or something that has had impact on the state of the match. All events have a timestamp and a position. Ball events are notated as 1 out of 19 tags, such as *Pass*, *Running with ball*, *Shot on target*, *High catch* etc. Depending on which tag is notated, the involved player(s), body part, cause, how, result and goal zone can be notated as well. For example, the notation of a body part is included when the ball touches the player when receiving a ball. This can be the *Left foot*, *the Right foot*, *the Header*, *the Chest* or *Two hands*. "How" implies on the situation such as *Running*, *Jump* or *Sliding*. "Goal zone" only applies to a *shot (not) on target*. Match events describe fouls, goals, free kicks, corners, cards and when the ball is out of play. They are often a result of a ball event.

Time	BallEvent	MatchEvent	Cause	How	P1	P2	X	Y	Result	WithWhat
6579	Running with ball			Running	11		256	301		Right foot
6608	Cross			Running	11		419	260		Right foot
6620	Clearance			Running	32		424	-17		Right foot
6664		Out for throw-in					298	350		
6692	Pass			Running	5		298	350	Throw-in	Two hands
6705	Reception			Running	7		323	241		Chest
6717	Running with ball			Running	7		354	277		Right foot

Table 3.3: Example of how events are notated

In total there are 12 827 events of which 11 995 are ball events and 832 are match events. These events are derived from a combination of the motion analysis and manual input from observers.

3.2.4 Errors

The collection methods have some catches and this needs to be taken into account when drawing conclusions.

- The height of the ball is not notated. When a long pass is given and it bypasses several players, we do not know if the ball flew 5 *m* above them or the players couldn't intercept the ball.
- The velocity of a player can be determined and when this velocity is high enough, it can be assumed that the player is facing the same direction. However when the velocity is lower, the player can be walking sideways or even backwards. It is thus not certain what the direction the player is facing.
- There is no accuracy known for the provided data. Visual inspection showed a player or the ball jumping from one place to another between two successive frames. This indicates that the positional data is not

always accurate.

- Besides cameras also observers notate events during a match. These notations contain errors and are open for interpretation.

3.3 Pass extraction

Eventually, the aim is to examine passes and find characteristics of "good" passes. In order to do so, the raw data needs to be processed in order to get a list of all passes which were successful. In order to be successful a pass needs to comply with the following rules:

- The ball event needs to be of the following types *Pass, Deep forward pass, Cross, Clearance* or *Neutral clearance*.
- The ball needs to reach a fellow teammate.
- The pass can not be a free kick which goes directly to the goal or a penalty, because they count as shot on target.
- A corner or free-kick which directly induces a shot on target should be excluded, because these passes result from a static time span in which all players could reposition themselves.

When a ball event fits all the rules, it is considered a successful pass. From such a pass, the following information is collected:

- The event itself.
- The n^{th} place in the current passing sequence, the so-called sequence number and the difference in time since the sequence started.
- The origin of the beginning of that sequence, which can be an *interception*, *throw-in*, *kick-off*, *corner* or *free kick*.
- The result after the sequence. This can be an *interception*, *out of play*, *foul*, *shot on target*, *shot not on target*.
- All positions in the time span of 1 second before the event occurs untill the pass is with the receiving player.

At the end a total of 3 378 so-called pass items remain and can be used for answering the research question.

Chapter 4

Feature construction

Several features need to be constructed in order to gain information about a pass. All these features can be classified into several categories: Fundamental, Context, Strategy, Opportunities. A pass itself has physical properties, such as length and speed. In addition, they originate from a certain state (context), and after the pass, this state has changed. These states and the differences might be a useful characteristic of a pass and will therefore be included. In total we constructed 21 types of features which result in a set of 72 features for each pass. The amount of features which could have been constructed are much higher. However we believe that these 72 features cover most aspects of the pass, since the types are accentuated from different views.

4.1 Fundamental

The features mentioned below describe the pass on a basic level, without including the context of the match.

Length X,Y ($\Delta x, \Delta y$) *(2 features)*
The length does not tell much about the direction of the ball in one particular way. Movement over the X-axis can be a defensive play backwards or a pass to the front of the field. For the Y-axis it can be a meaning of wide play. Therefore, Δx and Δy are features.

Length *(1 feature)*
The length of a pass is calculated by taking the start and end positions of the pass using Pythagorean triple. $length = \sqrt{dx^2 + dy^2}$.

Time *(1 feature)*
Time is calculated by calculation the time difference (Δt) between the start and end time stamps.

Speed *(1 feature)*
The speed in *km* per hour is constructed by dividing the Length (*meters*) by Time (*"second"*) and multiplying this by 3.6. This is because the

Direction *(2 features)*

We construct 2 different features for the angle of the direction of the pass. First, we have a single feature that describes the angle of the pass direction. The range is $\{x \in R | -180 > x <= 180\}$, where 0 degrees is when the pass goes towards the opposition her goal line parallel to the X-axis. Second, we have a categorical feature with three different categories: *forward, backward* and *sideways*, as shown in Figure 4.1.

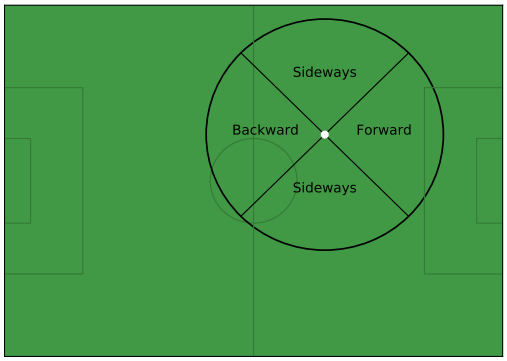

Figure 4.1: The way a pass gets categorised based on the angle of the direction the pass is going.

Body part *(1 feature)*

This feature is not constructed, but is already present in the event describing the pass. It can be a *left foot, right foot, header, chest* or *two hands*. This is a categorical feature. Note that *two hands* only applies when it is a throw-in or when the goalkeeper plays the ball.

Pass classification *(1 feature)*

Also this feature originates from the event. In Section 3.3, the events are considered a pass, if a certain event has type *Pass, Deep forward pass, Cross, Clearance* or *Neutral clearance*. This type is notated as a feature.

4.1.1 Surprise element

Some passes can be surprising and others can be predictive. The following features try to express the surprise element of a pass.

Pass angle *(2 features)*

The ball is carried by a player and she has a direction in which she moves. When a pass is sent to another player, the direction is not necessarily the same as this direction. It can also go sideways or even backwards. This can be a surprising move for the defenders and thus be an interesting feature to investigate. The angle is notated the same as with the direction feature.

Receiver distance to receive position *(1 feature)*

When the ball is played from one player to another, all players can move across the field. A pass to a receiver who did not move can be considered less surprising than when the receiver covers some distance in order to

receive the ball. The distance the receiver covered is the measure for this feature.

4.2 Context

When a player has the ball, the defending team can put pressure on her. Also the pass giver is at a position on the field. These are examples of local information of the context in which the pass is sent to another player. Besides, features such as the mean of the positions of each team also give context, but they tell more on the global state.

4.2.1 Local

Fieldposition *(4 features)*
The position of where the ball is passed gives context to the pass. A pass can originate from the back or from the middle of the field. The X and Y coordinates are both used as a feature. These two alone do not give much context. Therefore, the field is divided into zones using two approaches. The first way emphasises the position over the X-axis and separates the field into three equal parts, *Back*,*Middle* and *Front*. The other way is a six by three grid system, as seen in Figure 4.2. Taylor et al. [13] used this grid system for analysing scoring opportunities and concluded that possession in zone 14 is crucial for goal scoring. Both "zone" features are categorical.

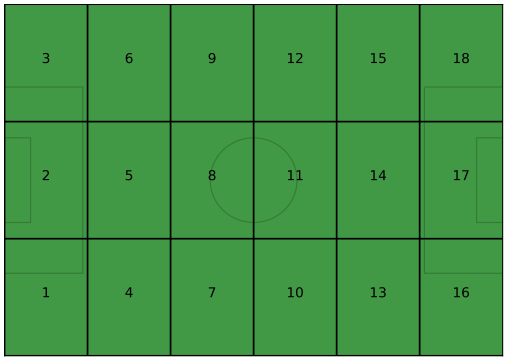

Figure 4.2: The positioning of the 18 zone grid over the pitch in order to categorise the positions.

Relative positions *(6 features)*
In addition to the position on the field, there is also a position relative to the rest of the players. If all players are on one half, an attacking player can pass from the centre line. Based on the absolute positions, she is in the middle of the field. However, compared to the rest of the players she is in the back. By calculating the centroids of a group of players, a relative position can be calculated. When the distance between this centroid and player is the measure for this feature, then a comparison with another pas cannot be made. Consider two situations where the player is 10 m away from the centroid. In the first situation all players are very close to

each other, whereas in the other situation they are more distant from each other. Relatively, the player in the first situation is in the back whereas in the other situation the player is more in the middle of the group. That is why the distance is divided by the mean distance of all players, in the group, to the centroid. This is done for a single axis or for the euclidean distance, which results in three different features for each group. Two different groups are used, namely the players of both teams together and the two teams seperately. This gives a total of six features in total. As an example, a negative centroid offset on the X-axis means that the player is in the back compared to the other players. Figure 4.3 represents the way the relative euclidean position of 1.83 is computed.

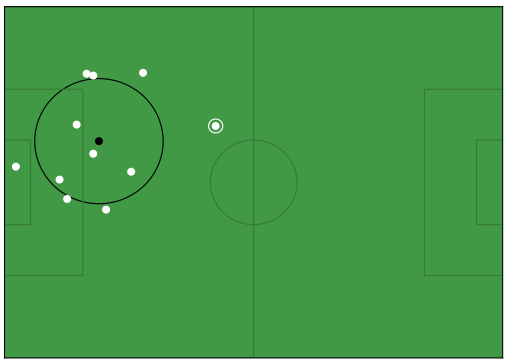

Figure 4.3: Representation of the the relative position feature. The white dots represent players, the carrier has a white circle around her, the black dot represents the centroid and the black circle is the density.

4.2.2 Pressure

InRange (pressure) *(24 features)*

A way to measure pressure is to count the number of opposing players within a certain radius around the player passing the ball. The higher the number of opposing players within a certain region, the more chance that a defender can intercept the ball. The radii which are used are 2 m, 4 m, 6 m and 10 m. The number of own players, opposite players and all players are separate features for each radius. This gives 12 different features for a given time. For the moments when the pass is given and received these features are constructed. This results in a total of 24 features.

Pressure zone (dangerousity) *(1 feature)*

To describe the pressure in more detail, we use the so-called pressure zone that is introduced in Ref [14]. This model assumes that a defender who is between the goal and the pass sender can apply more pressure than defenders at the sides and at the back. The presence of more defenders does not double the pressure, but increases the pressure logarithmically. There are four zones of which each have different radii (Figure 4.4) The pressure is expressed as a value between 0 and 1. Equation 4.1 is the individual pressure of a defender which takes the euclidean distance between the ball carrier and the defender (d_{D_i}) and the angle degrees the defender is of the carrier-goal line (α). $r_{ZO}(\alpha)$ gives the radius of the zone based on α. When a defender is in

the High Pressure Zone, the pressure is automatically equal to 1. In all other zones, the pressure increases as the defender comes closer. Equation 4.2 computes the total pressure and x is the sum of individual pressures for all defenders who are inside the Pressure Zone. The constant k influences the rate on how quick the total pressure increases based on x. For all passes in the data set, the highest value for x is 2. By selecting $k = \frac{e}{2}$ the pressure will remain within a range of $(0, 1)$.

$$PR_{D_i}(d_{D_i}, \alpha) = 1 - \frac{d_{D_i}}{r_{ZO}(\alpha)} \tag{4.1}$$

$$PR(x) = 1 - e^{-kx}, where\ x = \sum_{\forall\ D_i\ inside\ PZ} PR_{D_i} \tag{4.2}$$

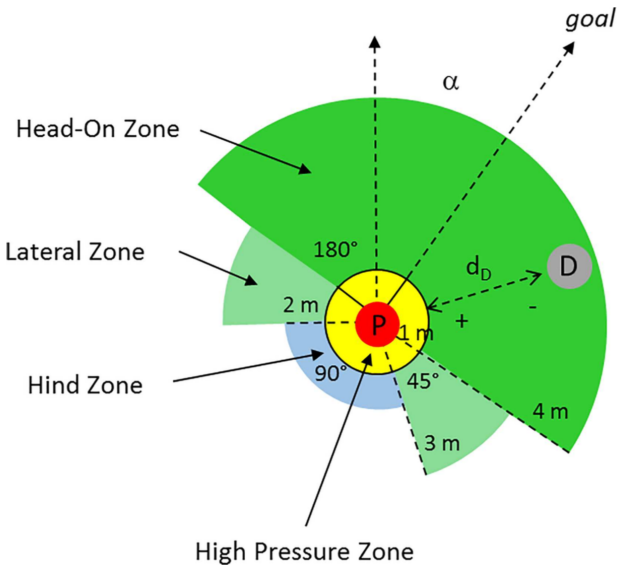

Figure 4.4: Geometry to determine Pressure. The Pressure Zone covers four areas with different radii, which result from the angle (α) between carrier and the goal. Pressure depends on the sub-area and the distance (d_D). Ref [14]

Area *(2 features)*

Kim et al. introduced Voronoi diagrams to soccer [15]. Voronoi diagrams divide the playing field into areas in which a certain player is dominant. The dominance region of a player is defined as the collection of points on the pitch that are closest to this player. These areas are called dominance regions and open up new possibilities to measure the context the match is in. The area of the dominance region of the pass giver gives insight into the pressure, but also the occupation of the regional space around her. It also gives a strategic insight. For example, when the area of the receiver is larger than the area of the pass giver, it can be assumed that the ball is in a more open position. The areas consist out of Delaunay triangles. By taking the sum of the area of all triangles, the total area can be calculated. The area of the sender and the receiver are computed and represent 2 features.

Adjacent players *(3 features)*

By examining the dominance regions, the number of adjacent players can be determined. This is done by counting the adjacent "friendly" and "hostile" regions as seen in Figure 4.5. By counting the number for

regions each type, 2 features are constructed. A high number of "friendly" regions could indicate there are a more pass possibilities. Where "hostile" players could indicate a higher pressure applied by the opposition.

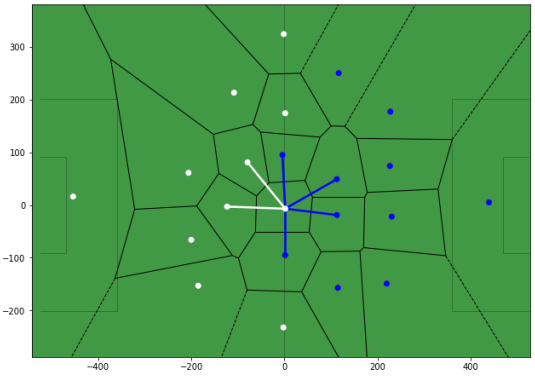

Figure 4.5: Dominance region with the adjacent players.

4.2.3 Global

Density *(4 features)*

The density is represented as the mean distance of position for each player to the centroid of the group. A group is a team or is both teams together. If this number is small, the players are more dense than when the number is higher. We compute the density for both teams separately and also combined. These represent three features. Also the ratio between the density of both teams can give context, since this tells something about the relative density to each other. This is the fourth feature.

Mean *(4 features)*

Over the X-axis a team has a certain mean position. In both directions, we can determine the mean position of all players of a single team or both teams together. These are three separate features. The mean tells us where the team is as a whole. This gives context, but also the difference between the means of the two teams do. For example, when the mean of the attacking team is higher than the defending team, then the pressure of the team overall, is relatively high. This is also a feature.

Sequence *(1 feature)*

A passing sequence is defined as the length of passes within one possession. This possession can end when the opposition intercepts the ball, the ball gets "out of play" or when a foul is made. Figure 4.6 shows a passing sequence with length 7 and is represented by its positions. Another representation can be seen in Figure 4.7 where the passes are plotted over time. It is assumed that when a pass is longer in roulation within a team the opposition tend to apply more pressure. The sequence is notated as a time stamp from the beginning of the sequence until the current pass. This gives more context than just the number of the position in the sequence.

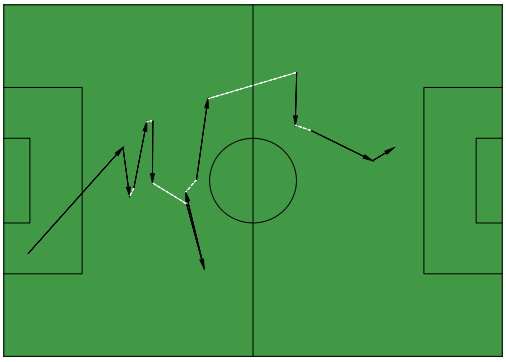

Figure 4.6: A passing sequence represented on the positions where the passes were given and received. A white dotted line is a dribble

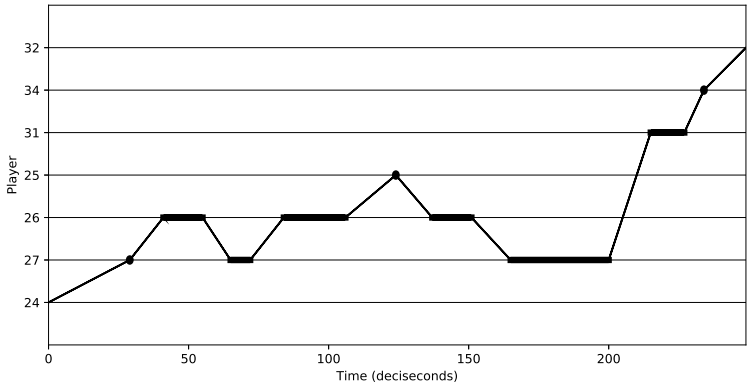

Figure 4.7: A passing sequence plotted over time

4.3 Strategic

A pass is given in a certain state, therefore a pass is given in a certain context. After the pass, there is another state and thus a new context. This difference between these two states, can result in a better strategic position. For example, the distance to the goal, the remaining amount of defenders between the ball and the goal, number of pass options, pressure and space to dribble. The difference between the origination state and the end state is important.

Changes in pressure, dominance region features, mean and density can show the strategic changes in the match states. In total there are 13 features.

Bypass *(3 features)*
Ensum et al. showed that there is a negative correlation between the number of defending players between the ball carrier and the goal and the probability to score [16]. With a pass, this number of defenders may change and thus affects the probability to score. Teams are analysed both separately and combined. In total, this results in three features. The difference in amount of players who are between the ball and the goal at the

beginning and the end of the pass, is the value of the bypass. A negative bypass value means that the pass increased the number of players between the ball and the goal.

People between ball and goal *(3 features)*
As an immediate result of the previous subsection, the count of people between the goal and ball carrier is used as a feature. This value is computed for both teams separate and combined, resulting in three features.

Distance to goal *(2 features)*
When the distance to the goal is less than the beginning of the pass, the probability to score increases, see Ref [17]. This feature is the difference in euclidean distance from the sender and the receiver. Apart from the difference, the absolute distance to the goal at the moment the pass is given is notated.

Chapter 5

Results

The main objective for a team is to win the match. Two main variables are underlying to this, namely the number of goals scored and the number of goals the opposite team scores. During the match one wants to maximise the probability to score a goal and to minimise the probability of the opposition to score a goal. A pass can be seen as a transition from one situation to another. For soccer there is no evaluation function which rates each state on the pitch at a certain time. We can, however, select passes from which is known it improves the situation. For example, a pass which brings the team in the position to score a goal can be considered a good pass. In this chapter, such passes are compared to all the other passes using subgroup discovery.

5.1 Setup

For the experiments, Cortana was used. Cortana has settings for the search strategy, quality measure and the inductive constraints. An overview of the parameters that were used in the experiments is displayed in Table 5.1. Since both binary and numerical target attributes were used, the quality measure is different for each kind of experiment.

Search strategy	
Strategy type	beam
Numeric strategy	best
Numeric operators	$\leq, \geq, =$
Inductive constraints	
Refinement depth	2
Minimum coverage	2
Maximum subgroups	∞
Maximum time	∞

Table 5.1: General parameter-setup

In order to find subgroups which are statistically significant from the entire dataset, the "measure minimum" was computed using a 100 times swap randomization. A 1% significance threshold is maintained during all

experiments.

5.2 Last pass before shot

The last pass before a *shot (not) on target* is made, was compared to all other passes. A shot is considered as a goal attempt and is considered a good situation. In principle, free kicks and corners also satisfy these requirements. However, both originate from a static situation and are therefore fundamentally different from the other situations. Therefore, corners and free kicks were excluded from the set. The set of passes was divided based on these conditions, which gave a binary target value of 116 "dangerous" and 3262 "other" passes.

Parameters	
Quality measure	WRAcc
Measure minimum	0.00569

Table 5.2: Parameters

Rules	**Coverage**	**Positives**	**Probability**	**Quality**
GoalDistanceStart $\leq$ 48.7016 AND GoalDistanceDiff $\leq$ 5.9487	634	76.0	0.1198	0.01605
GoalDistanceDiff $\leq$ -2.3806 AND GoalDistanceStart $\leq$ 65.8454	899	84.0	0.09343	0.01572
GoalDistanceStart $\leq$ 48.7016 AND Direction $\geq$ -135.0	734	78.0	0.1062	0.01562
GoalDistanceStart $\leq$ 48.7016 AND PositionY $\leq$ 28.5	694	76.0	0.1095	0.01544
GoalDistanceStart $\leq$ 48.7016 AND BypassOwn $\geq$ -3	761	78.0	0.1024	0.01535

Table 5.3: Top 5 subgroups

By using the parameters from Table 5.2 a total of 2041 subgroups were discovered. The 5 subgroups with the best quality are presented in Table 5.3. All subgroups had a relative high coverage and low probability. However, compared to the total set, this probability was much higher. The best subgroup translates to; "A pass which is given within a radius of 48.7 m from the goal and the pass brings the ball not more than 5.94 m away from the goal". The second subgroup is quite familiar to the first one, since the features are the same and only the values difference from each other. The next three all share the condition that the pass needs to be given within 48.7 m from the goal. The difference between these three subgroups is the additional condition. The third subgroup has the additional constraint that the direction needs to be ≥ -135 degrees. This implies that the ball can go everywhere except to the left-back. Next, the fourth subgroup has as a second constrained that a pass must not originate above 28.5 m on the Y-axis. The fifth subgroup has "$BypassOwn \geq -3$" as additional rule, which can be interpreted as that the ball may not increase the number of teammates between the ball and the goal by 4 players or more. These three additional constraints rule out a small portion of passes in addition of the first constraint.

Figure 5.1 presents the ROC space on which all the subgroups are plotted. A total of 4 subgroups define the ROC-curve and can be found in Table 5.4. The AUC is 0.772, which is fair.

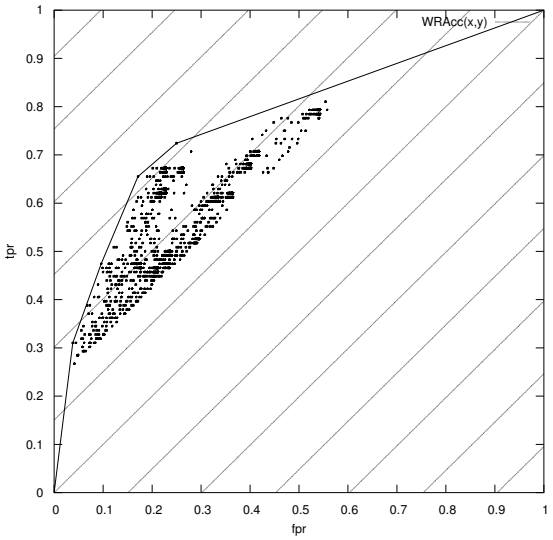

Figure 5.1: ROC space of the subgroups discoverd with the last pass before shot target. With an AUC of 0.772

FPR	TPR	Conditions
0.03801	0.3103	dY $\geq$ 17.3 AND GoalDistanceStart $\leq$ 52.1162
0.09596	0.4741	InRange100OppEnd $\geq$ 2 AND MeanOpp $\geq$ 21.6727
0.1711	0.6552	GoalDistanceStart $\leq$ 48.7016 AND GoalDistanceDiff $\leq$ 59.4873
0.2498	0.7241	GoalDistanceDiff $\leq$ -2.3806 AND GoalDistanceStart $\leq$ 65.8454

Table 5.4: Subgroup conditions defining the ROC-curve

5.3 Pass sequence before shot

As described in Chapter 4, a passing sequence is defined as passes which occur in sequence and are all within one team. Instead of looking at only the last pass before a shot is made, the whole passing sequence can be taken into consideration. By labelling all passes from passing sequences which led to a shot on goal as "good" and all the other as "bad", another target value is constructed. In total there are 341 passes which are in these passing sequences and 3026 which do not.

Parameters	
Quality measure	WRAcc
Measure minimum	0.012860119

Table 5.5: Parameters

Rules	Coverage	Positives	Probability	Quality
MeanOwn $\geq$ 3.3636 AND GoalDistanceDiff $\leq$ 7.8945	1136	202.0	0.1778	0.02475
GoalDistanceStart $\leq$ 52.1202 AND PositionY $\leq$ 34.2	951	182.0	0.1913	0.02454
GoalDistanceStart $\leq$ 52.1202 AND Direction $\geq$ -167.0	962	183.0	0.1902	0.02449
Direction $\geq$ -146.0 AND GoalDistanceStart $\leq$ 5.2120	917	178.0	0.1941	0.02441
GoalDistanceStart $\leq$ 52.1202 AND GoalDistanceDiff $\leq$ 7.8945	852	171.0	0.2007	0.02433

Table 5.6: Top 5 subgroups

By using the parameters from Table 5.5 a total of 743 subgroups were discovered. The 5 subgroups with

the best quality are presented in Table 5.6. The best subgroup translates to; "A pass where the mean of all own players is on the opposition his half and the pass brings the ball at most 7.8 *m* closer to the goal of the opposition". All the other subgroups have one shared condition: "GoalDistanceStart $\leq$ 52.1202", which is that the pass needs to be given within a radius of 51.2 *m* from the goal. Since the distance from the centre line to the goal line is 52.5 *m* this roughly means that the pass needs to be given on the half of the opposition. The second subgroup has an additional condition, the position over the Y-axis and can be interpreted as passes which originate not above 34.2 *m* on the Y-axis where the maximum is 38 *m*. The third and the fourth subgroup have almost the same additional condition which can be seen as "Passes which can go everywhere except to the left-back". The fifth subgroup relies on the distance to the goal and the same maximum distance rule as described with the first subgroup.

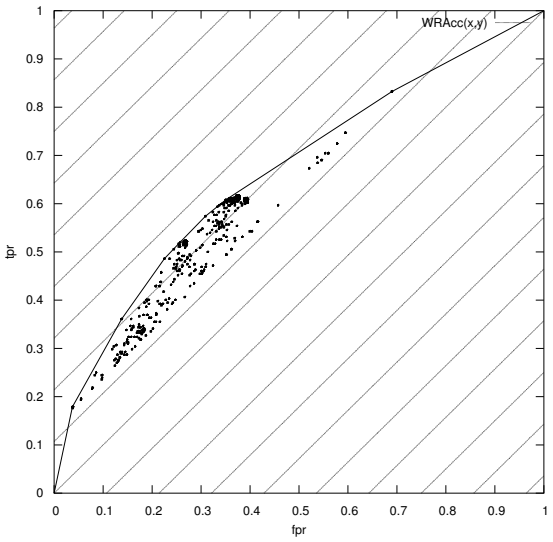

Figure 5.2: ROC space of the subgroups discovered with the last passingsequence before shot target. With an AUC of 0.665

Figure 5.2 presents the ROC space on which all the subgroups are plotted. A total of 6 subgroups define the ROC-curve and can be found in Table 5.7. The AUC is 0.665, which is poor.

FPR	**TPR**	**Rules**
0.03734	0.1790	dY $\geq$ 20.2 AND MeanOpp $\geq$ 12.5727
0.1378	0.3608	GoalDistanceStart $\leq$ 52.1202 AND DirectionCategorial = 'Sideways'
0.2250	0.4858	GoalDistanceStart $\leq$ 52.1202 AND GoalDistanceDiff $\leq$ 7.8945
0.2541	0.5170	GoalDistanceStart $\leq$ 52.1202 AND PositionY $\leq$ 34.2
0.3086	0.5738	MeanOwn $\geq$ 0.3364 AND GoalDistanceDiff $\leq$ 7.8945
0.3503	0.6080	MeanOwn $\geq$ 0.3364 AND How = 'Running'

Table 5.7: Subgroup conditions defining the ROC-curve

5.4 Pass sequence time regression

When looking at passes within a passing sequence, not all passes are of the same importance for bringing the team in a position to score. The last pass probably has more influence than a pass which occurred 12 seconds before the shot. By computing the target value based on the time difference between when the pass was given and when the actual shot was made, this can be taken into account. Therefore, we define the target as:

$$Target(t) = e^{-\frac{t}{\Delta t}} \tag{5.1}$$

Equation 5.1, where t is the time difference and Δt is the mean of all passing sequence durations. For our set-up this is equal to 12.6 second. The value of this target attribute decreases logarithmically over time. For all passes, the mean of the target value is 0.04897. In order to find subgroups, cWRAcc was used as a quality measure. Since it is a regression problem, an ROC-space cannot be constructed.

Using the parameters from Table 5.8 a total of 120 subgroups were discovered. The 5 subgroups with the best quality are presented in Table 5.9. There is one dominant condition for all these 5 subgroups, namely "GoalDistanceStart $\leq$ 48.73708", which means that the pass needs to be given within a radius of 48.7 m from the goal. The "Bypass" conditions mean that at least that amount of players need to be bypassed. "All", "Opp" and "All", refer to the own team, the opposite team and both teams respectively. These three conditions rule out passes which are played to the back. The other conditions are the dominance areas of the sender (Start) and the receiver (End). Most players have a median dominance region far below $95.6m^2$ and $112m^2$ when they pass the ball. Only the goalkeepers occasionally have such a large dominance area. Since the characteristics of a dangerous pass are researched in this project, it is obvious that the passes of the keeper are not of much importance.

Parameters	
Quality measure	cWRAcc
Measure minimum	0.022229975

Table 5.8: Parameters

Rules	**Coverage**	**Average**	**St. Dev**	**Quality**
GoalDistanceStart $\leq$ 48.7370 AND BypassAll $\geq$ -9	793	0.1222	0.2721	0.04096
GoalDistanceStart $\leq$ 48.7370 AND BypassOpp $\geq$ -7	811	0.1198	0.2697	0.0409
GoalDistanceStart $\leq$ 48.7370 AND BypassOwn $\geq$ -5	814	0.1194	0.2692	0.04079
GoalDistanceStart $\leq$ 48.7370 AND AreaEnd $\leq$ 95.6	822	0.1182	0.2682	0.04075
GoalDistanceStart $\leq$ 48.7370 AND AreaStart $\leq$ 112	824	0.1179	0.2680	0.04058

Table 5.9: Top 5 subgroups with the cWRAcc quality measure

Chapter 6

Conclusions

In this thesis, dangerous passes in soccer were investigated. From 6 matches of the WEURO 2017, a total of 3366 passes were selected which do not originate from a static situation. For each pass 72 features were constructed, consisting of 21 different types. The three approaches to label dangerous passes have been used. By using Subgroup Discovery, we have found that with these three approaches several conditions are dominant in the discovered subgroups. The best subgroups all cover a large part of the population, but show a significant difference in the proportion of dangerous passes. The characteristics are not only specific to dangerous passes, but hold a higher proportion of them. In answer to the research question, "What are characteristics of dangerous passes in soccer at the WEURO 2017?", we can conclude that dangerous passes occur in a higher proportion on the attacking half of the field (Distance to goal $\leq$ 48.7 m, 65.8 m or 52.1 m). Also the difference in this distance is an occurring constraint, which should be less or equal to 5.9 m or 7.8 m. Other constraints rule out very specific passes of the total population and are therefore not considered useful characteristics.

Although this research gives some interesting insights, there are certain restrictions. First of all, the results are quite biased to the strategies of the Dutch team. A larger, more generalised set of matches might give less biased results. Secondly, the collection methods give certain uncertainties on the found solutions, since the data is not 100% accurate. Thirdly, some features are constructed based on conclusions from other papers [13–17]. All these papers based their research on male soccer matches. Our matches are performed by women and although they are playing the same game, there is no guarantee that the conclusions in these papers also apply to women's soccer. Comparing womens to mens soccer might rule out or confirm this uncertainty.

6.1 Future research

Based on the research done in this project, a lot of future research can be conducted. Here, only 6 matches without a diverse test set were used as input. By increasing the number of matches and adding a more diverse distribution of teams, we can obtain results of higher generalisability. Another possible direction of future

research is considering other target attributes. We consider passes which have led to a shot as "good", but by looking at specific zones, possibly other, less trivial characteristics can be found. Except from using rules for labelling, other methods such a manual labelling can be used to identify characteristics of "good" passes. Besides looking at dangerous passes, the domain can be expanded to all passes which result in a higher scoring probability.

Bibliography

[1] Mike Hughes and Ian Franks. Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences*, 23(5):509–514, 2005. PMID: 16194998.

[2] Mike Hughes and Roger Bartlett. The use of performance indicators in performance analysis. *Journal of Sports Sciences*, 20(10):739–754, 2002. PMID: 12363292.

[3] Michael Horton, Joachim Gudmundsson, Sanjay Chawla, and Joël Estephan. Classification of passes in football matches using spatiotemporal data. *CoRR*, abs/1407.5093, 2014.

[4] Robert Rein, Dominik Raabe, and Daniel Memmert. which pass is better? novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science*, 55:172 – 181, 2017.

[5] IFAB. *Laws of the Game 2017/18*, 05 2017.

[6] Matthijs van Leeuwen and Arno Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 25(2):208–242, Sep 2012.

[7] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49.

[8] Nada Lavrač, Peter Flach, and Blaz Zupan. Rule evaluation measures: A unifying view. In *International Conference on Inductive Logic Programming*, pages 174–185. Springer, 1999.

[9] Ljupčo Todorovski, Peter Flach, and Nada Lavrač. Predictive performance of weighted relative accuracy. In Djamel A. Zighed, Jan Komorowski, and Jan Żytkow, editors, *Principles of Data Mining and Knowledge Discovery*, pages 255–264, 2000.

[10] Cortana subgroup discovery. http://datamining.liacs.nl/cortana.html [21-06-2018].

[11] *Regulations of the UEFA European Womens Championship*, 2015.

[12] Pascual J Figueroa, Neucimar J Leite, and Ricardo ML Barros. Tracking soccer players aiming their kinematical motion analysis. *Computer Vision and Image Understanding*, 101(2):122–135, 2006.

[13] A. Scoulding, N. James, and J. Taylor. Passing in the soccer world cup 2002. *International Journal of Performance Analysis in Sport*, 4(2):36–41, 2004.

[14] Daniel Link, Steffen Lang, and Philipp Seidenschwarz. Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLOS ONE*, 11(12):1–16, 12 2016.

[15] S Kim. Voronoi analysis of a soccer game. *Nonlinear Analysis: Modelling and Control*, 9(3):233–240, 2004.

[16] Jake Ensum, R Pollard, and Samuel Taylor. Applications of logistic regression to shots at goal in association football: Calculation of shot probabilities, quantification of factors and player/team. *Journal of Sports Sciences*, 22(6):500–520, 2004.

[17] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *Proc. 8th Annual MIT Sloan Sports Analytics Conference*, pages 1–9, 2014.