Applying Machine Learning to Event Data in Soccer

by

Matthew G. S. Kerr

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author . . . .
Department of Electrical Engineering and Computer Science
May 22, 2015

Certified by . . . .
John V. Guttag
Dugald C. Jackson Professor, Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Albert R. Meyer
Chairman, Master of Engineering Thesis Committee

Applying Machine Learning to Event Data in Soccer

by

Matthew G. S. Kerr

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2015, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

Soccer is the world's most popular sport but published research in soccer analytics has yet to attain the same level of sophistication as analytics in other professional sports. We hope to discover new knowledge from a soccer ball-event dataset by applying different machine learning techniques.

In this thesis we present three experiments that address three interesting questions in soccer that involve game prediction and team style. We approach each question by constructing features using the ball-event data, where an event is a pass, shot, etc., and applying machine learning algorithms. In the first experiment, we construct three models that use different features to predict which team won a given game, without any knowledge of goals. We achieve a top accuracy rate of 0.84 using an L2-regularized logistic regression classifier. We also investigate the feature weights to learn relationships between game events and a team's chances of success.

In the second experiment we try several classifiers to predict which team produced the sequence of ball-events that occurred during a game. Despite the relatively small number of events per game, we achieved an accuracy rate of 0.345 for a 20-team classification task when using a RBF SVM. By learning which sequences are characteristic of teams we are potentially able to discover if successful teams have a common style. We also learn the efficacy of transforming ball-events into predefined symbols.

Finally, in the third experiment, we predict which team attempted a given set of passes. We first construct 2D histograms of the locations of the origins of the passes. We then use the histograms as features in a 20-team classification task and discover that teams have characteristic passing styles, by achieving an accuracy rate of 0.735 using a learned K-NN classifier.

The results demonstrate that approaching soccer analytics with a machine learning framework is effective. In addition to achieving good classification performance, we are able to discover useful, potentially actionable, knowledge by investigating the models and features that we construct.

Thesis Supervisor: John V. Guttag
Title: Dugald C. Jackson Professor, Electrical Engineering and Computer Science

Acknowledgments

This thesis was supported by the help and advice of many. I would first like to thank the Qatar Computing Research Institute for their funding support and provision of the dataset.

I would also like to extend my sincerest gratitude to my thesis supervisor, Prof. John Guttag. His advice and ideas were invaluable and provided guidance whenever I was struggling. This thesis would not have been possible without his direction and support.

This thesis also benefited greatly from the generous help and advice offered by my lab mates. I would like to thank Guha Balakrishnan, Davis Blalock, Joel Brooks, Jen Gong, Yun Liu, Anima Singh, and Amy Zhao. Their feedback and critiques during group meetings inspired many ideas that are present in this thesis.

Finally, none of this would have been possible without my family and friends. My parents and brother have supported, guided, and helped me throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Soccer Analytics

In the past 15 years there have been significant efforts to further the field of sports analytics in both professional and academic settings. The publication of Moneyball [13], detailing the success of the analytical methods used by the Oakland Athletics baseball team and the rise of several prominent conferences dedicated to sports analytics has brought more attention to the field. This has led to increased amounts of sport data being collected, and an increase in the amount of analytics performed.

Although soccer is by far the world's most popular sport [4], published work in soccer analytics has yet to acheive the same level of sophistication as the analytics being performed in other professional sports. There is little use of data-driven analytics that are able to detect subtleties and insights that would otherwise be missed by human observation. Historically, much of soccer analytics has relied on developing new summary statistics to measure and confirm traditional beliefs about the game [3]. We believe that this type of statistical analysis does not fully leverage the newly available datasets in soccer, such as the dataset of ball-events we use in this thesis and describe in Chapter 2.

The work we present in this thesis uses machine learning techniques to learn new knowledge from a dataset of ball-events. We believe that our approach allows the data to "speak for itself." We do not begin with any prior beliefs or biases about how the game should be played. Instead, we first pose a question about an aspect of the game. We then construct features to represent the events that occurred during a match and aid us in addressing the initial question. We believe that the insights gained from our results will be able to help managers and players develop new strategies and gain a competitive edge. The experiments that we ran and present in this thesis are outlined in Section 1.2.

1.2 Experiments

In this thesis, in Chapters 3, 4 and 5, we present three separate experiments. For each experiment, we discuss our motivation, the problem we tackle, our methodology and the results and conclusions we obtain. The three problems we solve are:

1. Given the events of a game (excluding goals), can we predict which team won?
2. Given the sequence of ball-events, e.g., passes, shots, tackles, etc., of a game, can we determine which team produced the sequence?
3. Given a set of attempted passes, can we determine which team attempted the passes?

Problems 2 and 3 are quite similar, and fall under the more general goal of discovering the playing styles of teams. Problem 1, on the other hand, is related to the problem of game prediction, and investigates what aspects of a team's strategy are important for the team's success.

The methodologies we used for each experiment were fairly similar. In each case, we approached the problem from a machine learning perspective. Machine learning encompasses a range of techniques that focus on building models from example inputs to then make predictions or decisions [5]. For our experiments, we focus on supervised machine learning techniqes; in supervised learning, the model is learned from labeled examples. It is then used to make decisions on unseen data. Each example is represented by a set of features. Then, given the values of the feature set for an unlabled new example, and based on the labeled examples it has already seen, the model will make a decision and predict the label of the new example.

In the context of soccer, using a machine learning approach requires answering several questions. First, we must decide what constitutes an example, and what the corresponding labels are. For example, we can decide to treat a game as an example and its corresponding label would be the team that won. Another example would be to treat a possession as an example, and the label is whether or not the possession ended with a shot. This decision will affect the number of examples we are able to use. Then, we must choose the features. If we use an entire game as the example, the features could be a collection of statistics that describe the game. On the other hand, if a single possession is the example, the features could be the individual ball-events that occured during the possession. For each experiment, we describe what features we used, as well as the machine learning task that we solved.

1.3 Previous Work

Much of the published research in sports analytics, especially research that utilizes spatiotemporal data, has focused on sports that are easily discretized, such as baseball, American football and tennis [2]. These sports are easily broken up into individual at bats, plays or points, that have obvious, immediate outcomes, such as hits, yards gained or a won point. For example, a seminal work by Intille and Bobick used player tracking data to recognize different plays in American football [11]. It has proved to be much more difficult to perform similar work for more sports that are not as easily discretized, such as basketball and soccer. However, in both [12] and [17], the authors were able to utilize player tracking data for basketball to classify offensive plays and the movement patterns of offensive players. Similar work in soccer has proven to be even more difficult because possessions are more difficult to identify, and obvious outcomes, e.g., goals, occur very rarely and are often not correlated with a team's quality within a game. An example of this is the large number of upsets that occur during the FA cup, an English soccer tournament that frequently features matches between teams with glaring talent discrepancies.

As a consequence, until recently soccer analytics has focused much more on building probabilistic models to simulate game actions, and predict the outcomes of matches. Research by Reep and Benjamin developed models for the success rates of different length passing sequences [19]. More recently, in both [8] and [9], the authors attempted to predict the outcome of matches by using possession rates of different teams and other historical statistics, respectively, to develop probabilistic models. There has also been some work on the prediction of more short-term events, such as goals. In [18], the authors investigated the frequency of passing that occurs immediately before and after a goal has been scored. They found that in the 5 minutes preceding a goal, the team that scores plays a significantly greater frequency of successful passes than their average for the half, whereas the conceding team played significantly fewer accurate passes.

As the amount of spatiotemporal soccer data has increased, there has been more work that leverages the information the datasets provide. Spatiotemporal data allows analysts to learn the underlying mechanics of the game, such as style or player movements. Bloomfield *et al.* used player tracking information to investigate the physical demands of soccer and the work rates of different players [6]. In [10], the authors leveraged ball-event data and passing sequences to identify the playing styles of different teams. They described a passing sequence by the number of unique players involved in the sequence, and show that different teams use different sequences at different rates. Lucey *et al.* used ball-event data to infer the location of the ball throughout a game. Using this information, they constructed "entropy-maps" to characterize how different teams move the ball during a match [14]. In more recent work, the same authors used player tracking data to calculate the likelihood of a team scoring a goal in the next 10 seconds [15].

The work we present in this thesis similarly utilizes ball-event data to predict the outcomes of matches as well as discover the playing styles of teams. In each of the experiments we construct novel features that are not used in any of the related works mentioned above. We also use different machine learning algorithms to achieve our goals.

1.3.1 Related Sequence Classification Work

In experiment 2, which we present in Chapter 4, we attempt to classify sequences of ball-events. Sequence classification is most commonly approached by mapping sequences to points within a vector space, and applying traditional machine learning algorithms. This requires defining either a similarity measure, or a distance measure on the sequences. Some common choices are longest common subsequence, or the edit (Levenshtein) distance. In contrast, we treat different subsequences as words and reduce the original problem to a document classification task. This approach is motivated by the work presented in [7] and [20]. In particular, the approach presented in [20] for classifying time series is especially relevant, since it involves discretization of real-valued signals (that are analogous to our event locations), and is simple to implement.

1.4 Organization of the Thesis

The remainder of the thesis is organized as follows. We introduce the dataset that we used for all of the experiments we ran in Chapter 2. We then present the three experiments and discuss the individual methodologies and results in Chapters 3, 4 and 5 respectively. We end with a summary and our final conclusions in Chapter 6.

Chapter 2

Ball-Event Data

2.1 Overview

The dataset we use throughout all of the experiments presented in this thesis was collected by Opta [1]. It is the same type of data that is provided by Opta's F24 feed. The data are hand-labeled annotations of each ball-event that took place during the course of a match, for example, each pass, tackle, shot, etc. A ball-event is recorded any time a player makes a play on the ball, apart from dribbling. The full list of ball-event types is presented in Section 2.2. This data is collected for every match that occurs during a season of soccer. The dataset also includes additional information for each ball-event such as the location, the player involved, and the outcome. We call these additional variables 'ball-event descriptors' throughout the rest of this thesis. Some descriptors, such as location, time of event and player involved, are common for each ball-event type. Other descriptors, such as pass length, are unique to individual ball-events. Finally, there is another descriptor that is common for a subset of ball-event types: outcome. The descriptors for each ball-event type, as well as descriptions of the outcome descriptor, are presented in Section 2.2.

For the purposes of the experiments we present in this thesis, we assume that the data are clean and accurate. This assumption is mostly valid as the data was collected and verified multiple times by humans. However, there are some descriptors that are based on subjective, qualitative decisions. For example, the dataset includes a descriptor for each pass that is a label for whether or not the pass was a "through-ball". We ignored any such qualitative descriptors for all of the experiments to avoid obtaining results that are biased by any subjective decisions, and do not list them in Table 2.2.

Several of the figures presented in this thesis refer to teams by their team ID number in the dataset. Table 2.1 presents the ID number for each team in the figures we included.

Table 2.1: Team Names and Team IDs

Team ID	Team Name	Team ID	Team Name
174	Athletic Club	185	Real Betis
175	Atletico de Madrid	186	Real Madrid
176	Celta de Vigo	188	Real Sociedad
177	Espanyol	190	Real Zaragoza
178	Barcelona	191	Valencia
179	Sevilla	192	Real Valladolid
180	Deportivo de La Coruna	450	Osasuna
181	Mallorca	855	Levante
182	Malaga	1450	Getafe
184	Rayo Vallecano	5683	Granada CF

2.2 Ball-Events

In Table 2.2, we list each of the different ball-event types, as well as any associated descriptors. For the purposes of brevity, we do not list ball-events that we did not use. We also do not include any descriptors that are common to all ball-event types. These are: unique event ID, game ID, time of the event, location of the event, the player involved, and the team of the player involved. The location of the event is given as an (x, y) coordinate; since there is no standard pitch size, the dimensions of each pitch are normalized to 100 x 100.

As mentioned, several ball-events have the associated "outcome" descriptor. It is a binary descriptor that equals 1 if the ball-event was successful for the team performing the event. For example, if a pass reached its intended target, outcome equals 1. In Table 2.2, for any ball-event types that have the "outcome" descriptor, we include a brief description of its meaning. If the ball-event does not include the "outcome" descriptor, this implies that the inclusion of the ball-event indicates that it was successful.

For both the pass and shot ball-events, there are several associated descriptors that describe the type of pass or shot. For example, the "cross" descriptor for a pass indicates whether or not the pass was a cross into the box. The "other body part" descriptor indicates whether or not the player used a body part other than his foot or head to perform the shot. The descriptors are binary, and are listed in Table 2.2.

Table 2.2: Ball-Event Types and Associated Descriptors

Type	Descriptors	Outcome Description
Aerial Challenge	Outcome	Player won the aerial duel
Attacking Move	Type of Move	
Ball Recovery		
Block		
Challenge		
Clearance	Outcome	Clearance reached a teammate, or went out of play
Dispossessed		
Error		
Foul	Outcome	1 if player commited the foul, 0 if player was fouled
Interception		
Keeper Event		
Pass	Outcome, End Location, Distance, Angle, Intended Recipient, Cross, Header, Free Kick, Corner	Pass reached intended teammate
Shot	Final Location of Ball, Distance, Free Kick, Penalty, Header, Other Body Part, On Target, Goal	
Tackle	Outcome	Player gains possession
Take On	Outcome	Player successfully passed opponent
Touch		
Turnover		

Chapter 3

Experiment 1: Game Prediction

In this section, we present an experiment in which we attempt to predict the winner of a soccer match based on the events that occurred during the match, using three different models. More formally, the problem we solve in this experiment is: given the events of a game and the two teams that participated (*team A* and *team B*), did *team A* win? We ignored goals in our model and only looked at features that are not obviously correlated with the outcome of game, since this problem is easily solved if we have knowledge of all the goals that happened during the game.

We begin this section with a discussion on the motivating factors behind this experiment. Then, we describe the data that was used in the experiment, the different features constructed for each model and the model building process. We finish by discussing the results for each model and commenting on our findings.

3.1 Motivation

The type of ball-event data that we described in Section 2 was only recently made accessible and as such, there has been a relatively small amount of published research that utilizes such a data set. In order to discover the usefulness of the data, we tackled the question of what contributes to a team winning or losing. By constructing three different models — one model that only uses general, statistical features that are commonly given in match reports, another model that uses features that are specific to this type of dataset, and a final model that combines both sets of features — we explore how we can utilize this dataset to create features that aid us in answering non-trivial questions. We refer to the general, statistical features as “obvious” features, and the features we construct using this dataset as “non-obvious” features. We explain what the obvious and non-obvious features are in further detail in Section 3.2.1.

If we find that a model that only uses non-obvious features performs as well, if not better, than a model that uses obvious features, we can conclude that these features are useful for this task. We will also learn what non-obvious aspects of a soccer game are associated with the outcome of game. Although we are only able to learn what features are correlated with a team’s chances of winning, rather than what features are causal, it is a first step in developing an intution that will help guide strategic playing decisions in the future. Another benefit of our methodology is that it does not make any assumptions with regards to what features are associated with the outcome of game; it allows the data to ‘speak’ for itself without any prior beliefs. By starting with an unbiased perspective, and then learning what features are important, we create a basis for future research — research that we describe in latter chapters.

3.2 Methodology

In the following sections we discuss how we performed the experiment. They are:

1. Transform the ball-event data into a dataset of outcomes and feature values. As mentioned, we created three different models by creating three different feature sets. Each model uses their different features to predict an outcome. We will refer to each model as the Obvious Model, the Non-Obvious Model and the Mixed Model.

Then, for each model, we separately:

2. Split the dataset into an 80-20 training-test split.

3. Normalize the feature values.

4. Train an L2-regularized logistic regression model using 5-fold cross validation to select the optimum penalty factor.

5. Report results on the test set and rank the features by their absolute weights.

3.2.1 Step 1: Transformation of the Original Data & Construction of Features

We began by creating a dataset that is suitable for a supervised machine learning task. Each game in the dataset is an example with an outcome y and associated feature values x. For each game in the dataset, we randomly chose a team to be *team A*. The other team is thus *team B*. The outcome for that game was 1 if *team A* won and 0 otherwise. For the purposes of this experiment, we ignored any games that ended in a tie.

We then calculated the different feature values for each model using the game data. The features for the Obvious Model and Non-Obvious Model are described in Tables 3.1 and 3.2 respectively. The features for the Mixed Model are a combination of all the features from the Obvious Model and Non-Obvious Model. Each feature was calculated for both *team A* and *team B*. We also included the difference in the respective values for *team A* and *team B* for each feature. If we let n and m be the number of rows in the feature tables for the Obvious Model and Non-Obvious Model, respectively, the resulting feature dimensionality for each model is $3(n-1)$ and $3m$ respectively.

The features for the different models were chosen by using expert knowledge and by looking at the traditional statistics that are most often recorded for soccer games [16]. The features are also split into the three different models in order to show that we are able to learn what non-obvious features are associated with the outcome of a match. We are also able to learn if there is a difference in the information provided by the different sets of features. The features included in the Obvious Model are features that would be easy to construct even if the ball-event data was not available; they are summary statistics of the entire game. However, the features included in the

Non-Obvious Model are features that would be difficult to construct if the ball-event data were not available; they rely on the location and other descriptive information of individual ball-events.

Table 3.1: The features of the Obvious Model and their descriptions. The value of each feature is calculated for both *team A* and *team B*, as well as the difference in the value.

Name	Description
Home	1 if *team A* played at home, 0 otherwise (binary)
Number of Shots	The total number of shots a team took during the game (continuous)
Number of Shots on Target	The total number of shots on target a team took during the game (continous)
Possession	The percentage of total time a team had possession of the ball during the game (continous)

3.2.2 Steps 2 & 3: Splitting the Data & Normalizing the Feature Values

After creating a dataset of outcomes and feature values, we took a random 80-20 split to create a training set and a test set. We also normalized the feature values using z-score normalization. We calculated the mean and standard deviation of each feature in the training set, and then normalized all the data (training and test set) using the values taken from the training set. The formula for the z-score normalization is:

$$z = \frac{x - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation of the training values for a particular feature.

Table 3.2: The features of the Non-Obvious Model and their descriptions. The value of each feature is calculated for both *team A* and *team B*, as well as the difference in the value.

Name	Description
Number of Attempted Passes	The total number of attempted passes by a team
Pass Success Rate	The percentage of attempted passes that were successful
Pass Length	The average length of all the attempted passes by a team
Number of Crosses	The total number of made crosses by a team
Number of Attacking Passes	The total number of passes attempted by a team in the attacking third of the field
Attacking Pass Success Rate	The percentage of passes attempted in the attacking third that were successful
Attacking Pass Length	The average length of all the passes attempted in the attacking third
Shot Distance	The average distance from goal of all the shots a team took
Number of Well Positioned Shots	The number of shots attempted from the middle third of the pitch, in the attacking third of the pitch
X Position of Tackle	The average x value for the locations of all the tackles attempted by a team
Y Position of Tackle	The average y value for the locations of all the tackles attempted by a team
Tackle Success Rate	The percentage of all attempted tackles that were successful
Number of Attacking Tackles	The total number of tackles attempted in the attacking third
Attacking Tackle Success Rate	The percentage of tackles attempted in the attacking third that were successful
Number of Defensive Tackles	The total number of tackles attempted in the defensive third
Defensive Tackle Success Rate	The percentage of tackles attempted in the defensive third that were successful

3.2.3 Step 4: Training the Model & The Classification Task

The task at hand is a binary classification problem, in which we predict the outcome of a match based on the features we created in Step 1. In order to perform the task, we used an L2-regularized logistic regression model. If the model predicted a value greater than 0.5 based on the feature values for a game, we predicted that *team A* won the game.

We chose an L2-regularized logistic regression model in order to provide a measure of interpretability once the model was trained. We were not only interested in the results of the classification task; we were also interested in learning which aspects of the game determined whether or not a team wins. By choosing to use logistic regression, we are able to use the coefficient value for a feature as a proxy for its importance in determining a team's success. We trained the model on the training set and performed 5-fold cross validation that maximized accuracy in order to select c, the regularization factor. The search space for c was: $\{0.001, 0.01, 0.1, 0.5, 1, 1.5, 10\}$.

3.2.4 Step 5: Results

In Section 3.3, we present the accuracy attained on the test set using each of the three models. We also rank the features of each model by the absolute value of their weights in order to gain insight into which features were important in determining whether or not a team would win the game. We present the top ten features for each model in Section 3.3 as well. However, we must be careful not to interpret the ordering of features as an absolute ranking of their importance. If subsets of the features are correlated, their individual weights may not be completely indicative of their importance; the absolute values of their coefficients may be smaller than if we included only uncorrelated features. However, by using an L2 regularization as opposed to an L1 regularization, we attempted to mitigate the effects of feature correlation.

3.3 Results

Each model was trained and tested using all of the games in the dataset that did not end in a draw and finished within regulation time. In Table 3.3, we present the accuracy obtained on the test set for each model, as well as the value for c that was chosen.

Table 3.3: Accuracy Rates, Value of c and Number of Features

Model	Accuracy	Best c	No. of Features
Obvious Model	0.74	0.1	10
Non-Obvious Model	0.76	0.1	48
Mixed Model	0.84	0.01	58

From the results, we see that all of the models performed reasonably well in classifying which team won the game, with the Mixed Model performing the best. All of the models also have higher accuracy rates than the baseline method of always choosing the home team to win, which yields an accuracy rate of 0.63.

In Tables 3.4, 3.5, and 3.6, we list the five features that have the greatest absolute feature weights. We also present the relative feature weights for each feature, which is calculated as follows: $RFW = \frac{|weight|}{\sum_{i=1}^{n} |weight_i|}$, where n is the total number of features. I.e., we normalized each feature weight by the sum of all the feature weights.

Table 3.4: The top 5 features found for the Obvious Model ranked by the absolute value of their weights.

Rank	Feature	Weight	Relative Weight
1.	Difference in the number of shots on target	1.03	0.259
2.	Number of shots on target *team A*	0.816	0.204
3.	Number of shots on target *team B*	-0.803	0.201
4.	Difference in the number of shots	-0.341	0.0855
5.	Home	0.335	0.0840

From the rankings, it is clear that a team's ability to generate shots, especially shots that are on target, is a powerful indicator of their success. For both the Obvious and Mixed Models, the difference in the number of shots on target is weighted heaviest, which is not surprising as a team can only score a goal if they generate a shot on target.

Table 3.5: The top 5 features found for the Non-Obvious Model ranked by the absolute value of their weights.

Rank	Feature	Weight	Relative Weight
1.	Difference in the number of crosses	-0.522	0.0618
2.	Number of crosses *team A*	-0.445	0.0526
3.	Difference in the average shot distance	-0.428	0.0507
4.	Number of crosses *team B*	0.414	0.0489
5.	Number of well positioned shots	0.401	0.0474

Table 3.6: The top 5 features found for the Mixed Model ranked by the absolute value of their weights.

Rank	Feature	Weight	Relative Weight
1.	Difference in the number of shots on target	0.672	0.0929
2.	Number of shots on target *team A*	0.555	0.0766
3.	Number of shots on target *team B*	-0.499	0.0690
4.	Home	0.433	0.0598
5.	Difference in the number of crosses	-0.394	0.0544

However, it is interesting to note that the number of crosses for a team is important in the Non-Obvious Model. The top two features are the difference in the number of crosses between the teams and the number of crosses team A makes during the game. It is not immediately apparent why the number of crosses a team makes during the game is related to their ability to win, especially when this relationship is not a positive one. The values of the coefficients for both the difference in the number of crosses and the number of crosses team A makes are negative, indicating that the model will predict that a team will lose if they make more crosses during the game. We discuss this result in more depth in Section 3.4.2.

3.4 Conclusion and Comments

3.4.1 Learning Non-Obvious Features

As stated earlier in Section 3.1, one of the goals of this experiment was to show that we could learn what non-obvious features, constructed using the available ball-event data, are associated with a game's outcome. One way to show this is to compare the performance of a model that only uses non-obvious features with the performance of a model that only uses obvious features. Another method is to compare the performance of a model that combines both sets of features with the performance of a model that only uses one of the sets. In this experiment, we try both methods.

First, we compare the performances of the Obvious Model and the Non-Obvious Model. We see that both models performed equally well (74.0% and 75.6% accuracy respectively), suggesting that the non-obvious features have a comparable of predictive power in aggregate. We are able to learn what non-obvious features are associated with the outcome of a game, which may help when making strategic choices in the future.

Next, we compare the performances of the Obvious Model with the Mixed Model. The accuracy of the Mixed Model was 13.5% greater than the accuracy of the Obvious Model at 84.0%. This shows that adding the non-obvious features results in models with greater predictive power. It also suggests that the information added by the non-obvious features is complementary to the traditional statistics.

Both comparisons highlight the usefulness of our approach. The results show that the non-obvious features can either be used as the basis of useful models, or be used to add more power to existing models. We were able to learn what non-obvious features are associated with a game's outcome, and show that the information they provided is complementary to traditional statistics. An advantage of these features is that they provide more actionable information. The non-obvious features include more information about location, and other ball-event details that a strategist can use to his/her advantage. Although the size of the non-obvious feature set is greater, which increases the dimensionality of the model, it provides a deeper understanding of a game than the information provided by traditional statistics.

3.4.2 Crosses and Wins

An example of the deeper understanding provided by the more specific features is the fact that number of crosses a team makes during a game is negatively related with the team's chances of winning. As this was a surprising result we investigated deeper in order to discover why there is such a relationship.

Intuitively, we believe that a cross should be an indication of a successful attacking move. The presence of a cross indicates that a team was able to bring the ball deep into the opposition's territory, creating a chance for a teammate to head or kick the ball in from the cross. We proposed several hypotheses that attempted to explain why we found the opposite. One hypothesis is that an unsuccessful cross, i.e., a cross that did not end in a goal, leaves a team vulnerable to attack. The player making the cross is out of position to defend, and more often than not the defending team is able to regain position after a cross. Another hypothesis is that teams that are unable to pass the ball into a good position for a shot rely on crosses to generate shots. The shots generated from a cross may not be good opportunities, and therefore the team is unable to score goals.

In order to test the two hypotheses, we calculated both the correlation between the number of crosses a team makes during the game and the number of goals they score, as well as the correlation between the number of crosses and the number of goals a team concedes. We calculated the two correlations across all the games in the dataset. The correlation between the number of crosses and the number of goals conceded is -0.008, which suggests that there is no link between crossing the ball and being out of position defensively. The correlation between the number of crosses and the number of goals scored is -0.0813. This suggests that crossing the ball more frequently results in either fewer opportunities to score, or worse opportunities to score.

3.4.3 Future Work

In this section, we propose several possible experiments for future work based on our results and observations. The first is to explore team specific models. The model we constructed in this experiment described above was a global model. We were able to only learn what features are associated with winning in general. However, if we constructed the same model, but only for a single team at a time, we would be able to learn what features are associated with winning for that team. This information could then be used by opposing managers in attempt to limit an opponents chances of winning. Furthermore, we would be able to learn what features are associated with winning for successful teams. We could then develop some intuition into what areas of the game are important for continued success.

Another possible experiment is to divide our initial model into an offensive model and a defensive model. In the experiment we conducted, we only learned what features are associated with winning, but did not learn how they are associated. In soccer, there are two components to winning: a team needs to score more goals while conceding less. We propose conducting an experiment that uses the same features for two linear regression models; one model will try to predict the amount of goals scored by a team during the match, and the other will attempt to predict the amount of goals conceded by a team. By doing so we would be able to learn what features are important for the offense of a team and what features are important for the defense of a team.

Finally, we propose conducting an experiment that builds before-the-fact models. The features for such a model would be constructed using data from matches that occured before the match whose result is being predicted. We would be able to learn what information is useful for predicting outcomes before a match takes place, i.e., what aspects of a team's play is most likely to guarantee future success. The data used for constructing the features can either be taken from the match that immediately precedes the current match, or it can be taken from all the matches that have already taken place. However, we believe that this is a more difficult problem, as there are factors that contribute to a team's success, such as fatigue, injuries, personnel changes, weather conditions, etc., that would be ignored by such a model. Nevertheless, if the model is successful, we would be able to learn useful, predictive knowledge.

Chapter 4

Experiment 2: Team Classification using Event Sequences

For experiment 2, we switched our focus from predicting the outcome of a game to the task of recognizing a team based on their style of play. Specifically, given all the ball-events that one team performed during a game, we seek to predict the team that produced the events. We begin with a discussion of the motivation for this experiment, then describe the process of building the model. This involves defining an alphabet of symbols that represent different events, transforming the data into the appropriate form, constructing and pruning the feature space, and finally choosing and training a classifier. We also discuss how we choose the hyperparameters for each step in the process. Finally, we describe our results and final conclusions.

4.1 Motivation

This experiment was motivated by two main factors. First, we wished to discover if we could quantify style. Expert commentators and spectators often talk about the style of play teams employ, but in very qualitative terms. A playing style encompasses the types of passes the teams attempt, the formation in which the teams' players line up, and other strategic choices made by the coaches and managers. Often, the features of a playing style are visually identifiable. For example, one can easily see whether it is a possession based strategy or a counter-attack strategy. A famous example is the “tiki-taka” playing style of Barcelona in recent years. We wished to discover if we could capture a team’s style in a quantifiable manner by building features from the available ball-event data such that a classifier would be able to successfully distinguish different teams.

The second factor is the benefits offered by mapping events to symbols, which are described in the results of previous work completed by this research group in basketball analytics. In [12], the authors propose representing a basketball play as a series of previously defined symbols. For example, a play might consist of an initial “pick,” followed by a “cut,” and finishing with a “corner three.” The reasoning is that even though the different events may differ in their locations, players involved, or time of occurrence, the net effect on the game is likely the same if the overall “action” is the same. For example, two “picks” that differ in location by less than a meter would probably have the same effect on the game play and should be treated the same. Our goal was to show the usefulness of applying the same principle of defining different symbols that correspond to different actions in soccer. We discuss this idea and our process of defining the symbols in greater detail in Section 4.2.1.

4.2 Methodology

In the sections below we describe the experimental process. The basic steps are:

1. Create an alphabet of symbols that describe the different ball-events that occur over the course of a game.

2. Transform each game into a sequence of symbols drawn from the alphabet.

3. Split the dataset into a random, training-test split that is stratified by team.

4. Construct the examples for the classification task. In this experiment, all the ball-events that occurred during a game are split into two series, one for each team. An example is a series of ball-events for a single team, and its corre-

sponding label is the team. The series of ball-events is transformed into a series of symbols and the feature values are calculated for that example.

5. Choose and train a classifier. In each step described above, we need to tune hyperparameters. We also need to choose a model type from a set of possible classifiers. We perform 5-fold cross validation on the training set to both tune the hyperparameters and select the classifier type.

4.2.1 Step 1: Defining an Alphabet

A match of soccer in our dataset can be viewed as a time series of different ball-events. As discussed in Section 2 each ball-event has a number of different descriptors some of which are continuous in nature. Because of the continuous nature of some of the descriptors (e.g. location), the space of possible ball-events is nearly infinite. Any attempts to perform any type of classification task or prediction task based on a sequence of ball-events would therefore likely be ineffective.

Consequently we introduce the idea of an alphabet. An alphabet defines a set of possible symbols. Each distinct ball-event maps to a single symbol drawn from the alphabet. We create an alphabet by grouping ball-events that we believe should map to the same symbol using domain knowledge. Different ball-events are grouped together first by the type of the ball-event, and then by quantizing the values of different descriptors. After we have defined an alphabet, a sequence of ball-events X can be transformed into a sequence of symbols drawn from the alphabet A — i.e., we have defined a mapping $F : [x_1, \ldots, x_n] \to [s_1, \ldots, s_n], s_i \in A$. The symbols add invariance to values of continuous descriptors to facilitate analysis.

The level of quantization is determined by five hyperparameters. The value of each hyperparameter is taken from a predefined set, and expresses the number of levels to which various ball-event descriptors are quantized. For example, we could quantize the location of an event to two levels, indicating only if it is in the attacking or defending half of the field. As such, the hyperparameters determine how many possible symbols exist in the alphabet — the larger the value of each hyperparameter, the more fine-grained the symbols become and the larger the set of possible symbols. The five hyperparameters are listed and described in Table 4.1.

Table 4.1: Hyperparameters of the alphabet

Name	Description
Success	A binary hyperparameter. Determines whether or not we differentiate between successful and unsuccessful ball-events for ball-event types that have an outcome descriptor.
PlayingZones	The number of zones into which we divide the playing field. Possible values: $\{1, 2, 4, 8, 12, 16\}$.
PassParam	The number of discrete values we divide "pass length" and "pass angle" into. For example, if $PassParam = 2$, we quantize "pass length" into 2 groups, and "pass angle" into 2 groups. For "pass length", we only quantized passes of length less than 10m, and grouped all longer passes together. Possible values: $1, 2, 3, 4$.
PassTypes	The number of possible pass types: Headers, free kicks. Possible values: $1, 2, 3, 4$.
ShotTypes	The number of possible shot types: Headers, other body part. Possible values: $1, 2, 3$.

The reason for defining sets of possible values is that it allows us to intelligently engineer quantization for the different descriptors. For example, we could have quantized the location descriptor by splitting the playing field into a number of equal sized zones, effectively turning the field into a grid. However, as shown in Table 4.1, the *PlayingZones* hyperparameter only takes values from the set $\{1, 2, 4, 8, 12, 16\}$. Instead of spitting the field equally in a linear fashion, we split the field into zones that we believe provide the most additional information. For example, if $PlayingZones = 4$, instead of splitting the field into four zones of equal size, we split the field into a defensive zone and three attacking zones. This is illustrated in Figure 4-1. This design choice, in theory, allowed us to better capture the differences between events taking place in more important areas of the field. We believe that distinguishing passes that take place in different attacking zones of the field is more important than distinguishing passes that take place anywhere in the defensive half.[1]

A similar approach was used when deciding upon the set of possible values for

[1] Future work should test the efficacy of this feature engineering versus the direct, linear splitting approach.

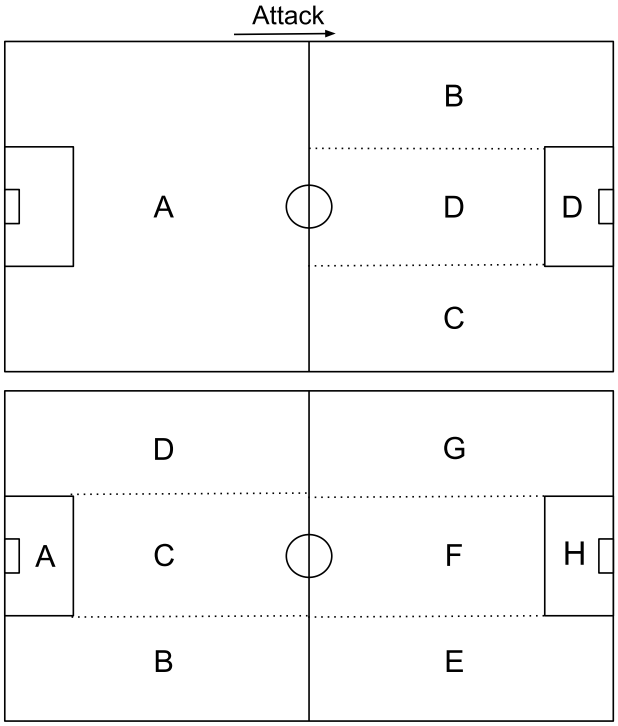


Figure 4-1: The different zones we split the field into, when $PlayingZones = 4$ and $PlayingZones = 8$.

the $PassParam$ hyperparameter. Figure 4-2 shows how we quantized the direction of a pass for different values of $PassParam$. We grouped "pass length" and "pass angle" together in order to reduce the space of possible parameter combinations, after initial experiments showed no significant difference in performance. The allowed values for the two hyperparameters discretizing the type of pass or shot, $PassTypes$ and $ShotTypes$, were constrained by the available, non-subjective binary descriptors for the respective ball-events. The descriptors for the Pass and Shot ball-events are discussed in greater detail in Section 2. Any descriptors that were based upon subjective human appraisals were ignored in our process.

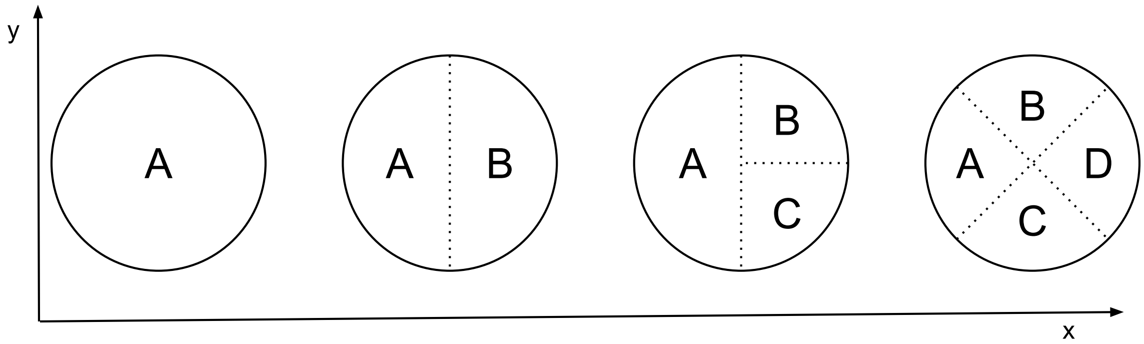


Figure 4-2: A diagram showing how we quantize the angle of a pass for different values of $PassParam$. From left to right, $PassParam = 1, 2, 3, 4$.

4.2.2 Step 2: Symbolizing

Once the values of the different hyperparameters are set and the alphabet is fully defined, the next step is to transform the series of ball-events into a series of symbols. We call this process "symbolizing." The first step of symbolizing a particular ball-event is to determine what type of ball-event it is. Then, depending on the type, we quantize the different relevant descriptor values using the constraints imposed by the hyperparameters of the alphabet. For example, a Pass ball-event will be quantized based on outcome, length, angle and location, while a Tackle ball-event will be quantized based only on outcome and location.

The size of the set of possible values thus depends on the number of different types of ball-events and the values of the hyperparameters. There is at least one symbol for each different type of ball-event, and the full size of the alphabet is equal to the total number of different combinations of discrete values a ball-event and its descriptors quantize to according to the values of the hyperparameters. We also include additional, unique symbols for penalty kicks, free kick shots, and corner kicks in our alphabet, even though these are not separate types of ball-events and their inclusion is not specified by any hyperparameter.

4.2.3 Steps 3 & 4: Splitting & Feature Construction

At this point in our experimental process, we have symbolized the series of ball-events of each team for each game in our dataset. Each series of symbols during a game for a team is thus an example in our learning problem, and its label is the team that performed that series. Therefore the number of examples in our experiment is $2n$, where n is the number of games in our dataset. After symbolizing every game, we randomly split the set of examples into a roughly 75-25 training-test split stratified by teams.

The features we construct are largely motivated by previous work done in sequence classification, since each series of symbols encoding a particular game can be viewed as a sequence. The previous work we examined was discussed in Section 1.3.1.

It should also be mentioned that we are primarily interested in the information offered by short subsequences (of length ≤ 4) for several reasons. First, we believe that the style of a team emerges in short sequences of action. For example, a team that depends on quick counter attacks may produce more subsequences that include two or three long passes followed by a shot, while a team that depends on possession may produce more subsequences that include five or more short passes before a shot ever occurs. Secondly, we are also interested in learning the impact different players have on games. We reason that this information would be more easily be extracted from short subsequences, since we would be able to associate the outcome of such subsequences with the players involved in order to learn their individual impacts.

Consequently, our first step was to extract all n-length subsequences from each sequence of symbols, where n is a small constant. These are analogous to "n-grams" within document classification and we henceforth refer to them as such. Given these n-grams, we then converted the set of all n-gram objects that occurred to counts of each subsequence within each game. To reduce the dimensionality of the resulting count vectors, and to ensure that the sequences used in classification were actually characteristic of a team, we then removed all subsequences from the lexicon that did not occur at least once per game (on average) for at least one team. That is, if no team typically produced a subsequence at least once per game, we dismissed that subsequence as noise. This is a bit *ad hoc* from a pure machine learning perspective, but since part of our goal was to show that sequences are characteristic of a team, extremely rare sequences would not be useful.

We then normalized the features using either a Term Frequency-Inverse Document Frequency (tf-idf) transform or a standard z-score normalization. The first emphasizes rare sequences, whereas the second simply forces uniformity in relative frequency and variability.

At this stage there often remained over a hundred features, something greater than the number of examples per team in our dataset. Therefore, we pruned the feature space further based on correlation with the output within the training data. Specifically, we selected the top $k\%$ of the features based on an ANOVA F-test.[2] Our choice of the ANOVA test over other statistical tests is somewhat arbitrary, and we could have similarly pruned the feature space based on mutual information, correlation, etc. Given much more computing power and time, we could have even carried out forward feature selection, or backwards feature elimination as part of our cross-validation, but this was impractical given the number of hyperparameters present within our symbolization procedure.

We could also have reduced the feature space through Principal or Independent Component Analysis but elected not to do so since it would have interfered with learning what subsequences are characteristic of teams. Knowing that a "corner kick, shot" is an especially common sequence is comprehensible; a linear combination of "corner kick, shot," "tackle, foul," and "pass, tackle" is not. Although we did not include matching teams with sequences in this experiment, we did seek to lay the foundation for it, and so we chose to ensure that our method produced results that were as interpretable as possible.

The final feature vector for each example is thus a vector of counts of different subsequences. This vector only includes counts of subsequences that appear often

[2] The exact parameters tried for k, as well as for other hyperparameters, are given in tables 4.2 and 4.1.

enough to pass our pruning techniques. The exact feature dimensionality depends on the final parameters chosen for each stage in our pipeline after the cross-validation process. The full pipeline is shown in Figure 4-3. For example, if we set $n = 2$, the size of our feature vector before pruning is m^2, where m is the number of different symbols generated by the alphabet. If every alphabet parameter is set to 1, m is equal to the number of ball-event types (17).

4.2.4 Step 5: Classification

After obtaining the transformed n-gram count vectors for each example, we classify the counts using one of a number of standard classifiers. In the case of linear classifiers (including logistic regression) we use a one-vs.-all approach, with the overall team label determined by the greatest distance to the separating hyperplane. The entire classification pipeline is shown in Figure 4-3.

For each step in our pipeline (except the removal of rare features) we needed to set the values of different hyperparameters. Furthermore, we also had to choose which classifier to finally use, and set the values of that classifier's parameters. The set of possible values for each hyperparameter is shown in Table 4.2. We carried out 5-fold cross-validation on the training set in order to choose the best value for each hyperparameter, as well as choose the best classifier.

We restricted the number of possible alphabet hyperparameters when carrying out the cross-validation. After preliminary testing with RBF SVM and Gaussian Naïve Bayes classifiers, we found that we attained peak performance when all such hyperparameters except for the PlayingZones parameter were set to 1. A more thorough investigation would have included the set of all possible values in the cross-validation process. However, this made the parameter search intractable.

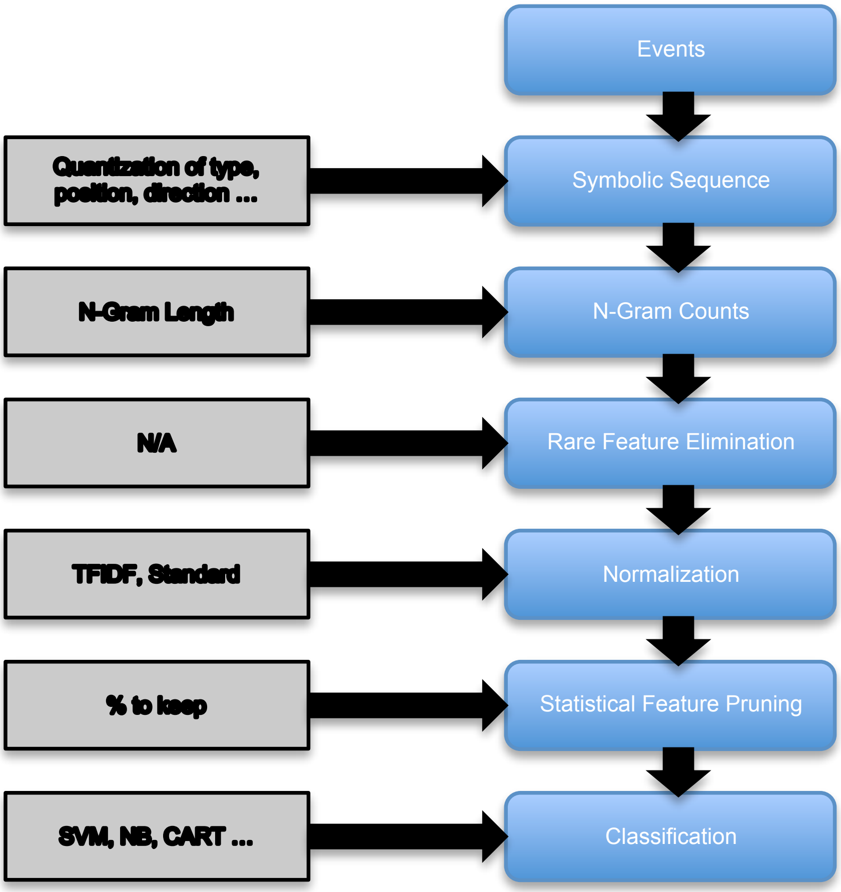


Figure 4-3: Pipeline used to classify games.

Table 4.2: Pipeline Parameters Tuned

Hyperparameter	Description	Values
	Alphabet	
(See Table 4.1)		
	Feature Extraction	
Ngram	Length of n-grams	1, 2, 3, 4
Normalization	Feature preprocessing	Standard (mean, variance), tf-idf
KeepPercent	% of features used	25, 50, 100
	Classification	
C	L2 regularization parameter for SVMs and LR	1, 10, 100, 1000

4.3 Results

In this section we present the results we obtained for the La Liga 2012-2013 season. This dataset contained 380 games which equates to 760 examples. 560 of these were used for training, and 200 for testing, which is roughly a 75-25 split. Since there are 20 teams in La Liga, the task we solve is a 20-way classification problem. In table 4.3, we present the best classification accuracies obtained for each classifier we tried on the test set, as well as the parameter values used to obtain peak performance.

Table 4.3: Best Classifier Performance and Associated Hyperparameters

Classifier	Acc.	Norm.	N-gram	Keep %	Zones
RBF SVM	0.345	Std.	2	100	16
Lin SVM	0.295	Std.	1	50	8
LR	0.315	Std.	1	50	8
NB	0.3	Std.	2	100	12
CART	0.175	Std.	1	25	16
AdaBoost	0.15	Std.	2	100	4

Despite the presence of twenty different teams, classifiers were able to achieve greater than 30% accuracy. For reference, if one were to classify teams at random, it would yield an accuracy rate of 5%. Since we tested 36 different classifiers, we applied a Bonferroni correction while testing for significance. The best result we obtained is significant at the 1×10^{-15} level. This decisively demonstrates that subsequences can be used to characterize different teams, and leads us to believe that style is a quantifiable property of teams.

We also observe that the optimal n-gram length is very short for all of the classifiers. There are two possible reasons for this phenomenon. The first is that the distinguishing characteristic of teams may be the number of times they perform certain ball-events in different parts of the field. This may also capture the skill of a team — better teams will probably perform more passes in the attacking zones of the field, for example. The second reason may be that longer subsequences (of three symbols or more) are too rare to effectively classify teams.

In order to better understand what errors the model was making we also plotted the confusion matrices of the two support vector machine classifiers in Figure 4-4.

We observe that both classifiers create similar looking confusion matrices. Both are able to classify Athletic Bilbao (team 174) and Barcelona (team 178) with very high recall. However, it also appears that there is a slight bias towards classifying other teams as these two teams. We see that the RBF SVM especially tends to classify Real Madrid (team 186) as Barcelona. The RBF SVM also does an especially poor job in classifying Espanyol (team 177) and Valencia (team 191). A final observation is that the Linear SVM also tended to make errors by wrongly classifying teams as Osasuna (team 450).

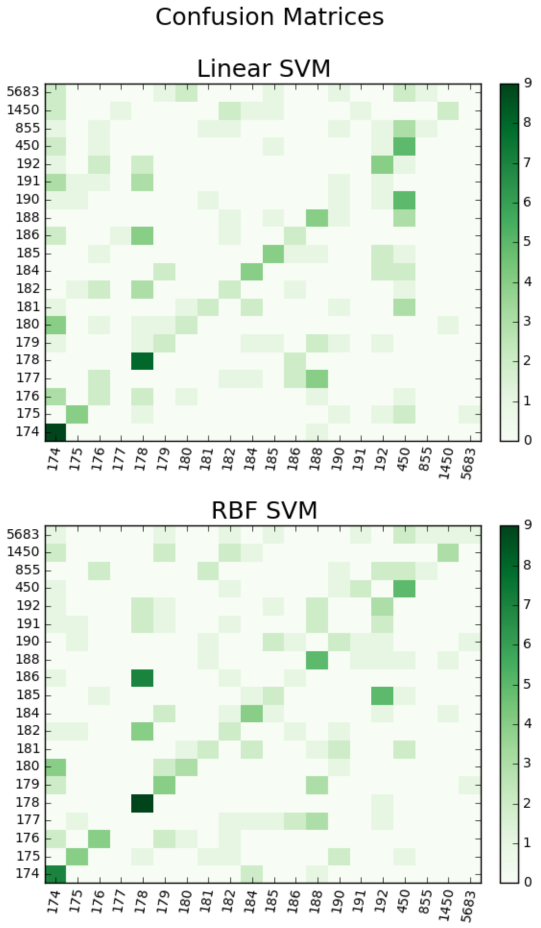


Figure 4-4: Confusion Matrices of different classifiers. Rows are true classes and columns are predicted classes.

There does not appear to be an underlying connection between a team's success and the model's ability to classify the team. Athletic Bilbao and Barcelona finished in the middle of the league and as champions, respectively. Real Madrid, Espanyol, Valencia and Osasuna all experienced different degrees of success as well. However, it should be noted that Athletic Bilbao and Barcelona produced the most ball-events during the season. It is unclear though, how this would affect the classification process. The count vectors are normalized and thus the raw number of ball-events should not affect the classifiers greatly. One hypothesis is that these two teams produce n-gram sequences that are not necessarily characteristic of their playing style often enough (through sheer numbers), such that the classifier mistakes other teams for Bilbao and Barcelona. For example, imagine that one of Real Madrid's characteristic sequences is (attacking pass, shot). This sequence may not be characteristic of Barcelona, but since Barcelona produces so many ball-events, they will also produce the same sequence during a game multiple times. This phenomenon may explain the bias we observe. We list the number of ball-events for each team, as well as the team's final position in the league, in Table 4.4.

Table 4.4: Number of Ball-Events and League Ranking

Team	# Ball Events	Position
Barcelona	72680	1
Atl. Bilbao	54644	12
Real Madrid	53540	2
Malaga	52844	6
Rayo Vallecano	52404	8
Real Valladolid	51782	14
Real Sociedad	51304	4
Sevilla	51278	9
Valencia	50560	5
Osasuna	49316	16
Deportivo	49228	19
Celta de Vigo	49138	17
Atl. de Madrid	49028	3
Real Betis	47348	7
Getafe	45838	10
Espanyol	45736	13
Real Zaragoza	45164	20
Granada	44422	15
Mallorca	43526	18
Levante	40408	11

4.4 Conclusion and Comments

4.4.1 Symbols versus Sequences

Based on the results shown in table 4.3, it appears that the optimal value of n is either 1 or 2. This suggests that sequences are perhaps not important when classifying teams. However, we believe that this would be the wrong conclusion to reach. For this experiment, after we set a value for n, we only include the counts of subsequences of length n in the feature vectors. This will bias n towards smaller numbers because of the rarity of longer subsequences, but we believe that this does not mean that subsequences of length 3 or more are not useful. For example, the sequence "pass, pass, pass" will probably occur far more frequently for Barcelona then it does for Levante, and will therefore be useful for team classification. A future experiment should include all the subsequences of length 1 through n before we prune our feature space. In this way, the feature vectors would include all of the useful subsequences of length $\leq n$, and not only those equal to n.

4.4.2 Quantifying Style

Our results show that it is possible to quantify style. Using reproducible, quantitative features, we were able to uniquely identify teams based on characteristic patterns. A team's style was reduced to numerical quantities. This result served as motivation for the experiment we present in Section 5.

4.4.3 Future Work

In order to expand on our findings, we propose several follow-up experiments. We previously mentioned in Section 4.4.1 we believe it would be useful to conduct the same experiment, except with counts of all subsequences of length $\leq n$ in the feature vectors, other than just those of a single length n. We believe that this would allow us to more accurately learn which subsequences are characteristic of different teams.

A different proposal is to explore what sequences are significant for individual teams. The work performed in the experiment above showed that there is reason to believe teams have characteristic styles that are reflected in ball-event sequences. Ideally, by learning what sequences differentiate teams, we would be able to discover if successful teams have similar characteristic sequences. This information could be used to make strategic choices and offer insight into the playing styles of opposing teams.

Another experiment would be to learn what sequences are associated with positive outcomes. We propose constructing a model that relates the features we used in the experiment described above with discrete outcomes that occur during a game, such as goals, goals allowed, penalty kicks, corner kicks, or winning the game. This information would provide insights into what aspects of a team's style are most important for success. Furthermore, if we build team-specific models, we would be able to learn what aspects of a particular team's style are most important. We would also be able to relate the sequences to the individual players involved and learn how different players impact the game.

Another similar experiment would be to analyze the variability in sequences within games and within teams across a season. The model we constructed makes the assumption that the sequences are drawn from a uniform distribution. However, this is often not true; fatigue and personnel changes will affect how teams play. By analyzing the variability within a game, we would learn how the playing styles of teams evolve during the course of a game. By analyzing the variability across games or halves, we would learn how consistent a team is with regards to their playing style. We hypothesize that more successful teams are able to play with a more consistent style since they have the ability to impose their style on a game. If we were able to discover the variability within different teams, we would be able to discover the relationship between consistency in style and success.

Chapter 5

Experiment 3: Team Classification using Passes

In experiment 3 we continue the theme of exploring the playing styles of different teams. However, in this experiment we focus only on the passes a team attempts and determine whether or not they are characteristic of a team. The formal definition of the problem we solve in this experiment is: given a random sample of passes made by a single team, which team attempted those passes? As in Chapters 3 and 4, we begin with a discussion of the motivation behind this experiment. We then proceed to outline our experimental process, and end by stating the results that we obtained and discussing the conclusions that we drew from this experiment.

5.1 Motivation

In Chapter 4 we explored the idea of creating an alphabet in order to define symbols to represent different groups of ball events. The results suggested that this idea had potential; the symbols that we created had useful information. However, the alphabet that we created in that experiment relied heavily on our intuition and domain knowledge. For example, when splitting the pitch into different zones, we chose the zones based on the areas of the field that we believed were most important. We believed that we could improve the usefulness of the alphabet if we created it in a data-driven manner.

One of the data-driven approaches we proposed was to cluster the ball-events based on a metric that captured the difference between two distinct ball-events. In order to define such a metric we needed to explore the distribution of ball-events. We specifically looked at pass events, since a pass was not only the most frequently occurring ball-event, but also had the greatest number of associated descriptors. We believed that there are three main attributes that differentiate passes: the length, angle and location. By plotting histograms of the values that appear in our dataset for these three descriptors, we were able to explore their respective distributions.

We constructed a 2D histogram of the origins of passes. In Figure 5-1, we present a histogram of all the passes that occurred during the 2012-2013 La Liga season. After constructing histograms using only the passes of individual teams, we noticed that they appeared to be visually very distinct, suggesting that a team's characteristic playing style was reflected in the origins of the passes they made. In order to confirm this observation, we investigated using the histograms as features in a team classification task. We discuss how these histograms are constructed in more depth in Section 5.2.2 below.

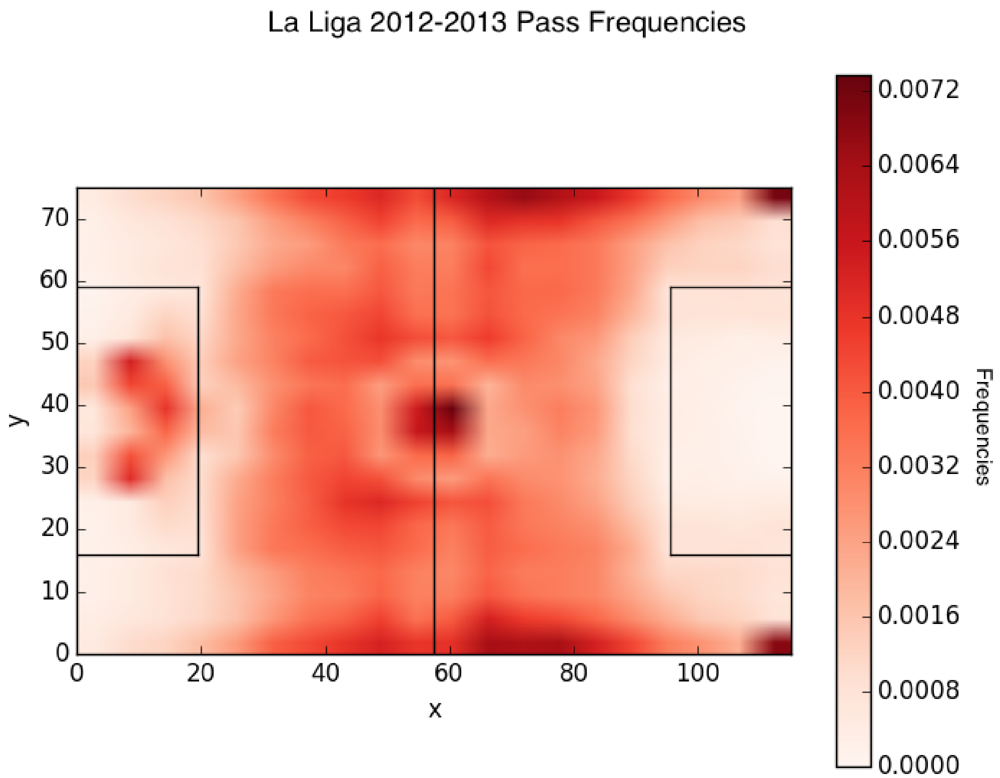


Figure 5-1: The 20 x 20 histogram of the locations of the origins of all the passes that were attempted during the 2012-2013 La Liga season. The histogram is normalized by the total number of passes.

5.2 Methodology

We describe our experimental process in more depth in the following sections. The individual steps of our process are described below:

1. For each team, randomly divide the set of all of the passes attempted by the team during the season into 10 subsets.
2. Construct a normalized, 2D histogram for each subset of passes.
3. Randomly split the data into a 70-30 training-test split.
4. Perform 3-fold cross validation on the training set to select the value of K for a K-NN classification model and use the resulting model to classify each

histogram of the test set.

5. Repeat steps 3-4 200[1] times and report the mean values of different metrics.

5.2.1 Step 1: Splitting the dataset

The first step in our experimental process is to split the dataset into smaller subsets. For each team in the dataset we collect every pass the team attempted (regardless of the outcome) to form a single set of passes. We then randomly assign each pass to one of 10 smaller subsets, creating 10 randomly constructed subsets of passes per team.

One of the alternatives to splitting the data in such a random manner would have been to split the dataset contiguously, i.e., create a subset of passes for a single game, or for a collection of sequential games. However in this experiment, we wished to explore the 'average' style of a team; we believed that a team's characteristic playing style is more identifiable when their play is looked at across an entire season. Teams will often change their tactics slightly depending on whom they are playing, introducing more variation. We attempted to abstract out this variation by randomly sampling across the entire season when constructing each subset of passes.

5.2.2 Step 2: Constructing the histograms

Each histogram represents a heat map of the origins of a set of passes. For a given set of passes, we divided the field into a 20 x 20 grid and count how many of the passes originated from each grid box. 20 x 20 was chosen arbitrarily; it prevented the matrix from becoming too sparse and still showed good performance during classification. We then normalized the histogram values by dividing the count of each grid box by the total number of passes in that set. In Figure 5-2, we show a histogram that was constructed using all of the passes attempted by Real Madrid during the 2012-2013 La Liga season.

[1] We chose 200 in order to obtain every possible training-test split with high probability.

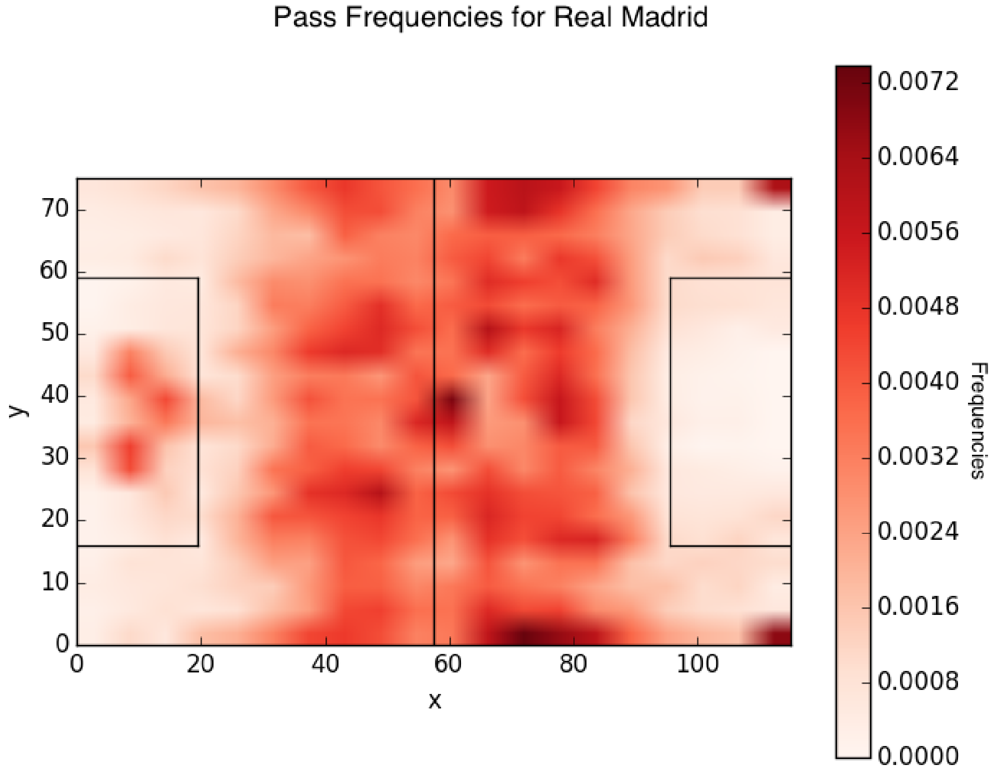


Figure 5-2: The 20 x 20 histogram of the locations of the origins of all the passes attempted by Real Madrid during the 2012-2013 La Liga season. The histogram is normalized by the total number of passes.

For the classification task, we treat each subset of passes as an example, and its respective label is the team that produced those passes. In order to use the histograms as features in the proposed classifier, we flatten the 20 x 20 histogram into a single vector of values. This creates a model that has feature vectors of dimension size 400, which is relatively large given that in our final experiment we only have 200 total examples (10 examples per team) in our dataset.

5.2.3 Steps 3 & 4: Classifying

After constructing a histogram for each subset of passes for each team we take a 70-30 training-test split stratified by team, thus ensuring the class balance is equal in both the training set and test set. We then construct a K-Nearest Neighbors classifier to perform the classification task. We chose a K-NN classifier because of the interpretability it offers, and also because it showed strong performance during preliminary experiments. In order to choose a value for K, we performed 3-fold cross validation on the training set and select a value for K from the set $(2, 3, 4, 5, 6, 7)$ — we chose the value that resulted in the highest mean accuracy across the folds.[2] The classifier also uses a weighted voting scheme; when classifying a single example, each of the K nearest neighbors will have a weighted vote proportional to the inverse of its Euclidean distance from the example.

After choosing a value for K, we classify each example in the test set and calculate the overall accuracy, as well as the precision and recall for each class. We present the results from our experiment below in Section 5.3. The values presented are the means of each metric after repeating steps 3 and 4 200 times.

5.3 Results

In this section, we present the results that we obtained when performing the experiment on the 2012-2013 La Liga season. There are 20 teams competing in La Liga, which thus transforms the experiment into a 20-way classification task. Each team on average made $\sim 18,000$ passes during the entire season, with Barcelona making the most passes by a wide margin at 30283, and Levante attempting the smallest number of passes at 13,094. Figure 5-3 shows the distribution of the number of passes each team attempted during the season.

After repeating the experiment 200 times as described in Section 5.2.3, we obtained a mean accuracy of 0.735 on the test set, with a standard deviation of 0.039. The mean value of K for each of the 200 classifiers was 5.78, with a standard deviation of 1.3. We calculated accuracy as the percentage of all the examples that we correctly classified. We also calculated the mean precision and recall values for each team, and

[2] The range of values we select K from is limited by the total number of examples for a single class in the training set, which is 7.

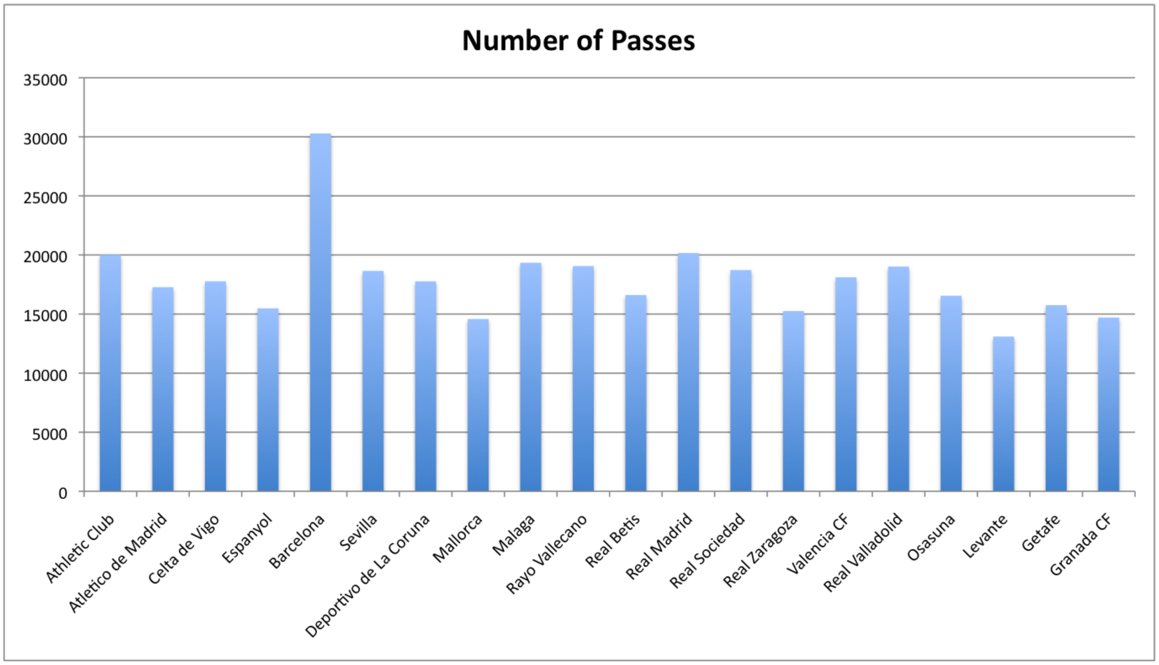


Figure 5-3: The total number of passes attempted by each team during the 2012-2013 La Liga season.

present them in Table 5.1. The precision of a given class C is calculated as: $p = \frac{t}{t+f}$, where t is the number of true positives, and f is the number of false positives. The recall of a class is calculated as: $r = \frac{t}{t+n}$, where t is still the number of true positives and n is the number of false negatives. In Figure 5-4, we plot the mean F-score for each team. We calculated the F-score as the harmonic mean of the precision and recall values: $F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$.

Table 5.1: Mean Precision and Recall Values for each Team, with Standard Deviations

Team	Mean Precision, Std. Dev.	Mean Recall, Std. Dev.
Athletic Club	0.965, 0.0867	1, 0
Atletico de Madrid	0.957, 0.095	0.963, 0.114
Celta de Vigo	0.658, 0.162	1, 0
Espanyol	0.166, 0.365	0.0617, 0.138
Barcelona	0.830, 0.153	1, 0
Sevilla	0.901, 0.137	0.977, 0.085
Deportivo de La Coruna	0.773, 0.168	0.987, 0.0653
Mallorca	0.490, 0.500	0.185, 0.205
Malaga	0.436, 0.140	0.960, 0.108
Rayo Vallecano	0.603, 0.164	0.992, 0.052
Real Betis	0.917, 0.244	0.647, 0.278
Real Madrid	0.947, 0.153	0.750, 0.244
Real Sociedad	0.764, 0.177	0.952, 0.117
Real Zaragoza	0.700, 0.458	0.290, 0.229
Valencia CF	0.782, 0.192	0.937, 0.131
Real Valladolid	0.624, 0.175	1, 0
Osasuna	0.870, 0.291	0.528, 0.267
Levante	0.480, 0.500	0.178, 0.200
Getafe	0.904, 0.227	0.702, 0.288
Granada CF	0.940, 0.218	0.590, 0.266

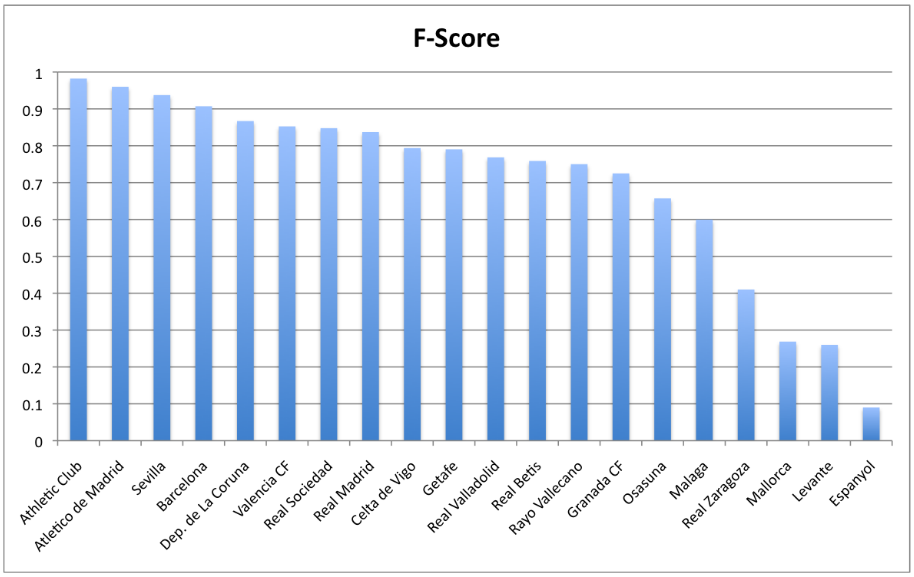


Figure 5-4: The mean F-score value for each team that participated in the 2012-2013 La Liga season.

5.4 Conclusion and Comments

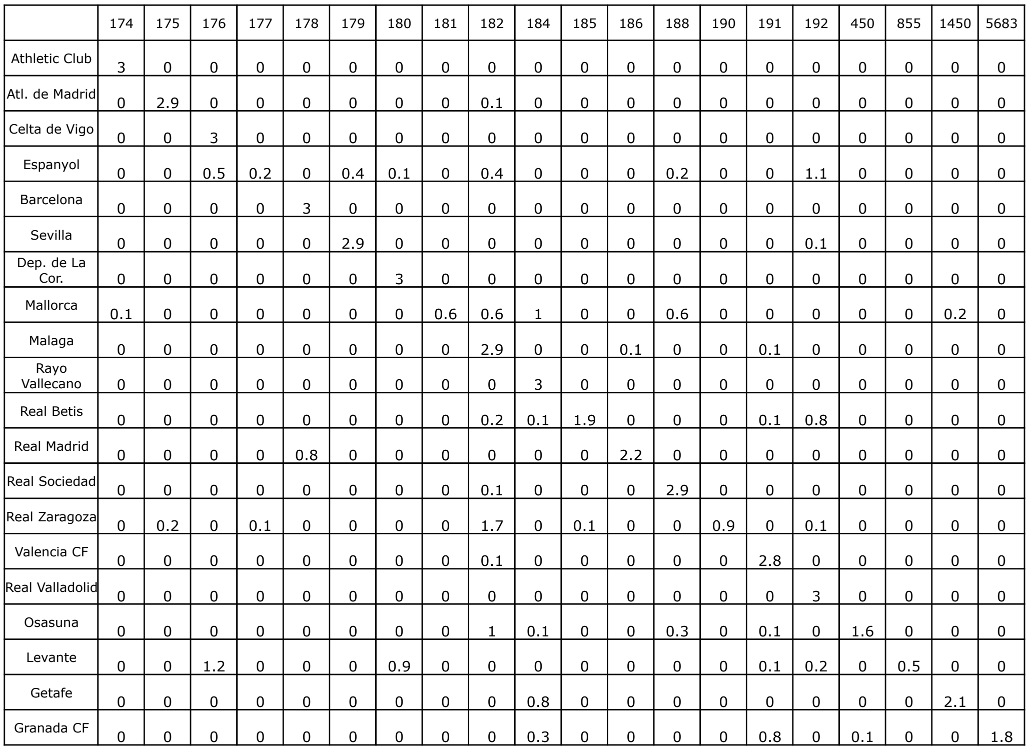
5.4.1 Performance

As shown by an accuracy rate of 73.5% we are able to classify teams with relatively good performance especially when we consider that choosing teams at random would only yield an accuracy rate of 5%. The accuracy rate we achieve is significantly better than random at a miniscule significance level ($<< 10^{-15}$). This suggests that the locations from which teams attempt passes are characteristic of a team and the 2D histograms are able to capture the different styles effectively.

This level of performance is not expected, especially when compared with the results of experiment 2. Although it is difficult to compare the two experiments — in experiment 2 we only observed the events of a single game, whereas in this experiment we look at passes from across the entire season — the difference in performance is still remarkable. It is not obvious why passing events should be more characteristic of teams than chains of ball events.

In order to better understand what kind of errors the model makes we plotted a confusion matrix (using the test set) as shown in Figure 5-5. The values of the confusion matrix are the mean values after repeating the experiment 200 times. We observe that for several teams we are able to correctly classify all instances of that class. This is reflected in the number of teams that have perfect recall, as shown in Table 5.1. However, the model underperformed for four teams in particular: Real Zaragoza, Espanyol, Mallorca, and Levante. These teams showed particularly poor precision and recall.

We inspected the data for these four teams in more detail in order to discover if there is an underlying reason why the model performed poorly with regards to these teams specifically. First, there appears to be a small correlation between playing success and poor classification performance. All four teams finished the season in the bottom half of the table. Second, we observed that the four teams attempted relatively few passes throughout the season. In fact, the teams were in the bottom five in terms of attempted passes, with Levante and Mallorca attempting the fewest



	174	175	176	177	178	179	180	181	182	184	185	186	188	190	191	192	450	855	1450	5683
Athletic Club	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Atl. de Madrid	0	2.9	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0
Celta de Vigo	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Espanyol	0	0	0.5	0.2	0	0.4	0.1	0	0.4	0	0	0	0.2	0	0	1.1	0	0	0	0
Barcelona	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sevilla	0	0	0	0	0	2.9	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0
Dep. de La Cor.	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Mallorca	0.1	0	0	0	0	0	0	0.6	0.6	1	0	0	0.6	0	0	0	0	0	0.2	0
Malaga	0	0	0	0	0	0	0	0	2.9	0	0	0.1	0	0	0.1	0	0	0	0	0
Rayo Vallecano	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Real Betis	0	0	0	0	0	0	0	0	0.2	0.1	1.9	0	0	0	0.1	0.8	0	0	0	0
Real Madrid	0	0	0	0	0.8	0	0	0	0	0	0	2.2	0	0	0	0	0	0	0	0
Real Sociedad	0	0	0	0	0	0	0	0	0.1	0	0	0	2.9	0	0	0	0	0	0	0
Real Zaragoza	0	0.2	0	0.1	0	0	0	0	1.7	0	0.1	0	0	0.9	0	0.1	0	0	0	0
Valencia CF	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	2.8	0	0	0	0	0
Real Valladolid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Osasuna	0	0	0	0	0	0	0	0	1	0.1	0	0	0.3	0	0.1	0	1.6	0	0	0
Levante	0	0	1.2	0	0	0	0.9	0	0	0	0	0	0	0	0.1	0.2	0	0.5	0	0
Getafe	0	0	0	0	0	0	0	0	0	0.8	0	0	0	0	0	0	0	0	2.1	0
Granada CF	0	0	0	0	0	0	0	0	0	0.3	0	0	0	0	0.8	0	0.1	0	0	1.8

Figure 5-5: The mean confusion matrix of the classifier. Each column represents the instances of the predicted class, and each row represents the instances of the actual class. The first row are the team IDs used in our dataset. The IDs are in the same order as the listed teams in the first column.

number of passes out of all the teams in La Liga that season. However, if we look at the confusion matrix in Figure 5-5, the model does not appear to make a common, biased mistake. The model mistakes the examples of Real Zaragoza, Levante, Mallorca and Espanyol for a variety of different teams.

In contrast, the model consistently misclassifies examples of some other teams, such as Getafe or Real Madrid, as examples of one other team only. For Real Madrid, the model would incorrectly classify on average 0.8 out of the 3 Real Madrid examples as Barcelona examples. Real Madrid is also the only team that has its examples misclassified as Barcelona. Curiously, Barcelona examples are never misclassified as Real Madrid examples. This suggests that Barcelona plays more consistently and that the histograms they produce are more homogenous, whereas the distances between

Real Madrid histograms are more variable. It also suggests that there is a similarity between the passing styles of the two teams, which may be a product of either similar coaching styles and/or similar skill levels of the teams' players. Nevertheless, the model is still able to classify Real Madrid examples with high performance, as shown by an F-score of 0.84.

5.4.2 Visualizing Connectivity

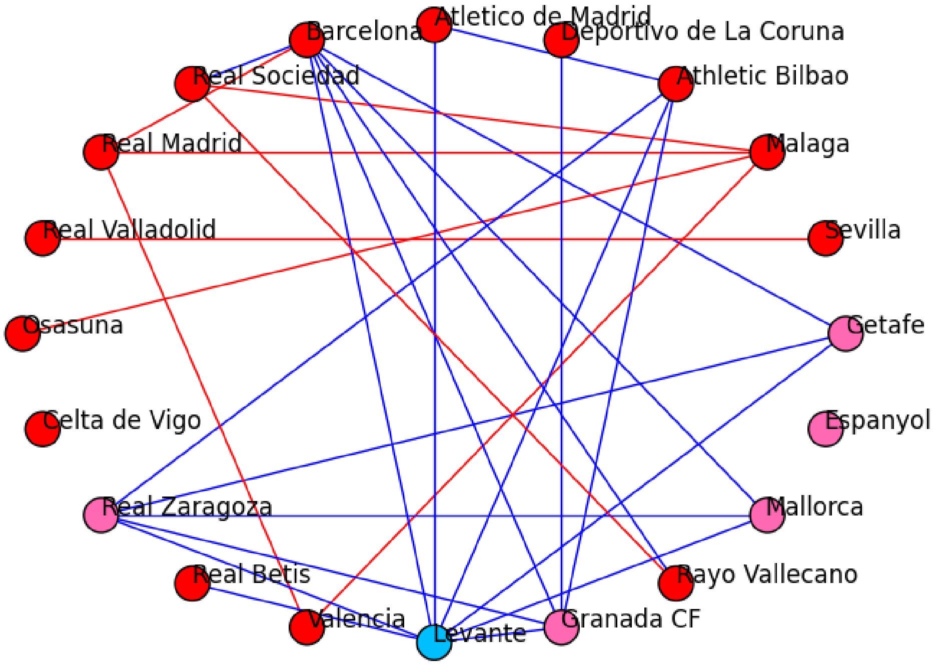


Figure 5-6: A visualization of the euclidean distances between pass histograms of teams. Each node represents a team. A red node means that the distance between histograms of the same team is within the first decile of all the distances. Similarly, a red edge means that the distance between the histograms of the two teams is within the first decile. A pink node means that the distance is between the tenth and fiftieth percentile. A light blue node means that the distance is between the fiftieth and ninetieth percentile. Finally, a blue edge means that the distance is within the last decile of all the distances.

In order to visualize the differences between the pass histograms of teams, we plotted the graph structure seen in Figure 5-6. The structure is constructed using a distance matrix, M, that we create after building the 2D histograms for each subset of passes. Each i, j entry, where $i \neq j$, is equal to the average distance between all the different pairs of histograms for a team, i, and a team, j. A mathematical formulation is:

$$M_{i,j} = \frac{1}{|H_i||H_j|} \sum_{(h_x,h_y) \in (H_i,H_j)} \|h_x - h_y\|$$

where H_i is the set of all the histograms constructed for a team i. Recall that we split the set of all passes attempted during a season for a team i into 10 random subsets; there are thus 10 histograms for each team and so $|H_i| = 10, \forall i$. Each i, j entry, where $i = j$, is simply the average distance between all the histograms constructed for team i. We repeat splitting the set of all passes, calculating the histograms for each subset and creating a distance matrix 10 times, such that the entries of the final distance matrix are averages over the 10 repetitions.

After creating the distance matrix, M, we are able to create a graph, G, in which each node, i, represents a team, and each edge, (i, j), represents the average distance between the histograms of team i and team j, i.e., the (i, j) entry of M. This graph is fully connected, and if we were to plot it as is, it would be indecipherable. Instead we only plot edges that represent extreme distances, and color code them according to their respective distances. In the example plotted below in Figure 5-6, we only included edges that represent distances that are within the first decile of all the distances in matrix M, or that are within the last decile. A red edge means that the edge represents a distance within the first decile, and a blue edge means that the edge represents a distance within the last decile. The nodes are also color coded according to the average distance between histograms of the same team. These distances correspond to the (i, i) entries of M. A red node means that the average distance is within the first decile, a pink node means that the distance is between the tenth and fiftieth percentile, and a light blue node means that the average distance is between the fiftieth and ninetieth percentile. I.e., red indicates relatively small distances, and blue indicates relatively large distances.

We are able to make several observations from the graph in Figure 5-6. The first is that there are several teams that appear to be relatively ‘distant’ from the rest. These are the teams whose nodes have a lot of blue edges emanating from them. These include Barcelona, Levante, Real Zaragoza, Granada and Athletic Club (labeled in the figure as Athletic Bilbao). This may imply that these are teams are quite distinct from others, and as such the model should perform well when classifying examples belonging to these teams. However, the model had various degrees of success with classification when measured by the precision and recall for the five teams. Athletic Club and Barcelona both have very high F-scores, 0.98 and 0.91 respectively. The rest of the teams however have relatively poor precision and recall scores — in fact, Levante has the second worst f-score of all the teams. Levante is also the only team that is represented by a light blue node, which may help explain the model’s extremely poor performance in classifying examples belonging to Levante.

Another observation is that the color of the edges emanating from the same node appear to be fairly uniform. It is very rare that a node is incident to edges of different colors. Only three teams show this behaviour: Barcelona, Rayo Vallecano and Real Sociedad. In addition, there are several nodes that have either very few or no incident edges. This suggests that the passing style teams use fall into three categories: there are teams that play quite similarly to several other teams, teams that have neither particularly similar nor distinct styles, and finally teams that play very distinctly to other teams. This may help explain the varying degrees of success our model has when classifying examples of different teams.

Finally, there are the teams that are represented by either pink or light blue nodes. This implies that the passing styles of these teams are inconsistent across a season. The teams are: Getafe, Espanyol, Mallorca, Granada CF, Levante and Real Zaragoza. Four of these teams had the lowest F-scores in our model, which we already mentioned in Section 5.4.1. After looking at the number of passes each team attempted, which is plotted in Figure 5-3, we noticed that these teams also attempted the fewest passes out of all the teams. Furthermore, they all finished in the bottom half of the final standings; Mallorca and Real Zaragoza finished in the bottom three. This suggests that these teams did not have enough possession of the ball, or perhaps enough skill, to play with a consistent style across the entire season. There is also the possibility that they simply did not have enough possession to even attempt enough passes such that they are able to play with any kind of consistent style. The lack of consistency is reflected by the lack of similarity among the histograms that represent their playing styles.

5.4.3 Future Work

In this subsection we propose several follow-up experiments to build upon our findings. A first proposal is to perform the same experiment, except we will split the set of all the passes attempted by a team contiguously instead of randomly. In Section 5.2.1, we discussed how we constructed 10 randomly sampled subsets of passes from across the entire season, in order to capture a team's general playing style. By splitting the set of all passes contiguously (perhaps by individual games or by sets of consecutive games), we will capture a team's playing style against individual opponents. We will be able to observe if teams are able to play with the same passing style against different opponents and test our hypothesis that more successful teams are able to play more consistently. If this hypothesis holds, we would observe that we are able to classify successful teams with greater accuracy.

Another proposal is to perform a similar experiment as the one described in this chapter, but using histograms that are constructed using the locations of different ball-events, as opposed to just passes. We believe that a team's overall playing style is captured by all the different ball-events and, as such, we would be able to achieve greater classification performance by combining all the ball-events. We would also potentially be able to learn how the different ball-events contribute to a team's playing style.

The histograms we constructed in the experiment captured the location of all the passes a team attempted regardless of its outcome or the possession's outcome. We propose constructing histograms of either all the passes that were attempted during a possession that led to a shot or only the first pass of such a possession. By doing so, we would explore how different teams generate 'good' possessions and from where the possessions originated. This would potentially allow us to develop an intuition into what kind of passes are 'good' and are beneficial to a team's offense.

Finally, we propose exploring if there are groups of teams that have similar passing styles. We have already shown in this experiment that there are underlying differences between teams that allow us to separate them. We also attempted to visualize how the 'distances' between different teams in Section 5.4.2. A next step would be to explore if there are groups of related teams that play with a similar passing style, by using clustering techniques. We would be able to learn if there are similarities amongst successful teams (or unsuccessful teams) which would be useful knowledge for strategists and managers.

Chapter 6

Final Summary and Conclusions

In this thesis we explored how we can approach soccer analytics from a machine learning perspective. We presented three experiments that explored different problems in soccer; the first explored predicting the result of games, whereas the last two attempted to discover the playing styles of teams. We approached all experiments with the same framework. We posed a question about an aspect of the game and then formulated the problem as a supervised machine learning task. For each task we used the same ball-event dataset to construct features. This approach allowed us to analyze soccer without any prior beliefs or biases.

The overarching goal of this thesis was to show that we can learn useful knowledge about soccer using machine learning techniques. The results of our experiments proved that we can accomplish different tasks with good performance. These results indicate that it would be fruitful to continue with a similar approach. Furthermore, we were able to develop new insights into the game of soccer using the features we constructed for each task. For example, in the first experiment we learned that there is a negative relationship between the number of crosses a team attempts during a game and their chances of winning. In the third experiment, we showed that we can distinguish different teams using histograms of their attempted passes. We believe that these insights, in combination with the future work we suggested for each experiment, will provide managers and players with a competitive edge when developing new strategies.

Bibliography

[1] OptaPro.

[2] SlamTracker.

[3] Soccer Analytics — Presented by Prozone | MIT Sloan Sports Analytics Conference.

[4] And the silver goes to... *The Economist*, September 2011.

[5] Machine learning, May 2015. Page Version ID: 662453525.

[6] Jonathan Bloomfield, Gudberg K. Jonsson, Remco Polman, Kenneth Houlahan, and Peter O'Donoghue. 16. temporal pattern analysis and its applicability in soccer. 2005.

[7] BYM Cheng and J G Carbonell. Protein classification based on text document classification techniques. *Proteins: Structure*, 2005.

[8] Christian Collet. The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010. *Journal of Sports Sciences*, 31(2):123–136, 2013.

[9] Anthony C. Constantinou, Norman E. Fenton, and Martin Neil. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36:322–339, December 2012.

[10] Laszlo Gyarmati, Haewoon Kwak, and Pablo Rodriguez. Searching for a unique style in soccer. *arXiv:1409.0308 [physics]*, September 2014. arXiv: 1409.0308.

[11] Stephen S. Intille and Aaron F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 518–525, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence.

[12] Mitchell Kates. Player motion analysis: Automatically classifying nba plays. Master's thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, sep 2014.

[13] Michael Lewis. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company, New York, 1st edition edition, March 2004.

[14] Patrick Lucey, Alina Bialkowski, Peter Carr, Eric Foote, and Iain Matthews. Characterizing multi-agent team behavior from partial team tracings: Evidence from the english premier league. In *AAAI*, 2012.

[15] 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.* MIT Sloan Sports Analytics Conference, 2014.

[16] Phil McNulty. Arsenal 0-0 Chelsea.

[17] Matej Perše, Matej Kristan, Stanislav Kovačič, Goran Vučkovič, and Janez Perš. A trajectory-based analysis of coordinated team activity in a basketball game. *Computer Vision and Image Understanding*, 113(5):612–621, 2009.

[18] Athalie Redwood-Brown. Passing patterns before and after goal scoring in FA premier league soccer. *International Journal of Performance Analysis in Sport*, 8(3):172–182, November 2008.

[19] C. Reep and B. Benjamin. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, 131(4):581–585, January 1968.

[20] P Senin and S Malinchik. SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model. *2013 IEEE International Conference on Data Mining (ICDM)*, pages 1175–1180, 2013.