THE UNIVERSITY OF SYDNEY

Classification of Passes in Football

Michael Horton

Joachim Gudmundsson, Sanjay Chawla

School of Information Technologies

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

INTRODUCTION

A knowledgeable observer of a game of football (soccer) can make a subjective evaluation of the quality of passes made between players during the game. We investigate the problem of producing an automated system to make the same evaluation for passes. We present a model that constructs numerical predictor variables from spatiotemporal match data using functions based on methods from computational geometry, and learns a classification function from labelled examples.

The importance that the learning algorithms assign to the predictor variables are analysed to determine if there is a relationship between the complexity of the algorithm that computed the predictor variable and the importance of the variable to the classification function.

MOTIVATION

There is considerable research into developing objective methods of analysing sports, including football. Football analysis has practical applications in player evaluation for coaching and scouting; development of game and campaign strategies; and also to enhance the viewing experience of televised matches. Currently, football analysis is typically done manually or using simple frequency analysis. There would thus appear to be scope to improve the efficiency of the analysis process as well as the quality of the output.

Our contribution is to explore the use complex methods from computational geometry to compute predictor variables and to then apply machine learning algorithms to train classifier functions. We will then run experiments on these classifiers to produce quantitative measures of passing performance.

PROBLEM DEFINITION

We consider the problem of classifying passes made during a football match based on the quality of the pass. The approach that we selected was to use supervised machine learning algorithms to learn a classification function. Such algorithms accept data of a particular structure, and thus a significant component of our approach is to take the raw spatiotemporal data and transform it to an appropriate format for the learning algorithms.

We used data from four matches played in the English Premiership during the 2007/08 season involving Arsenal Football Club. The data was provided by Prozone.

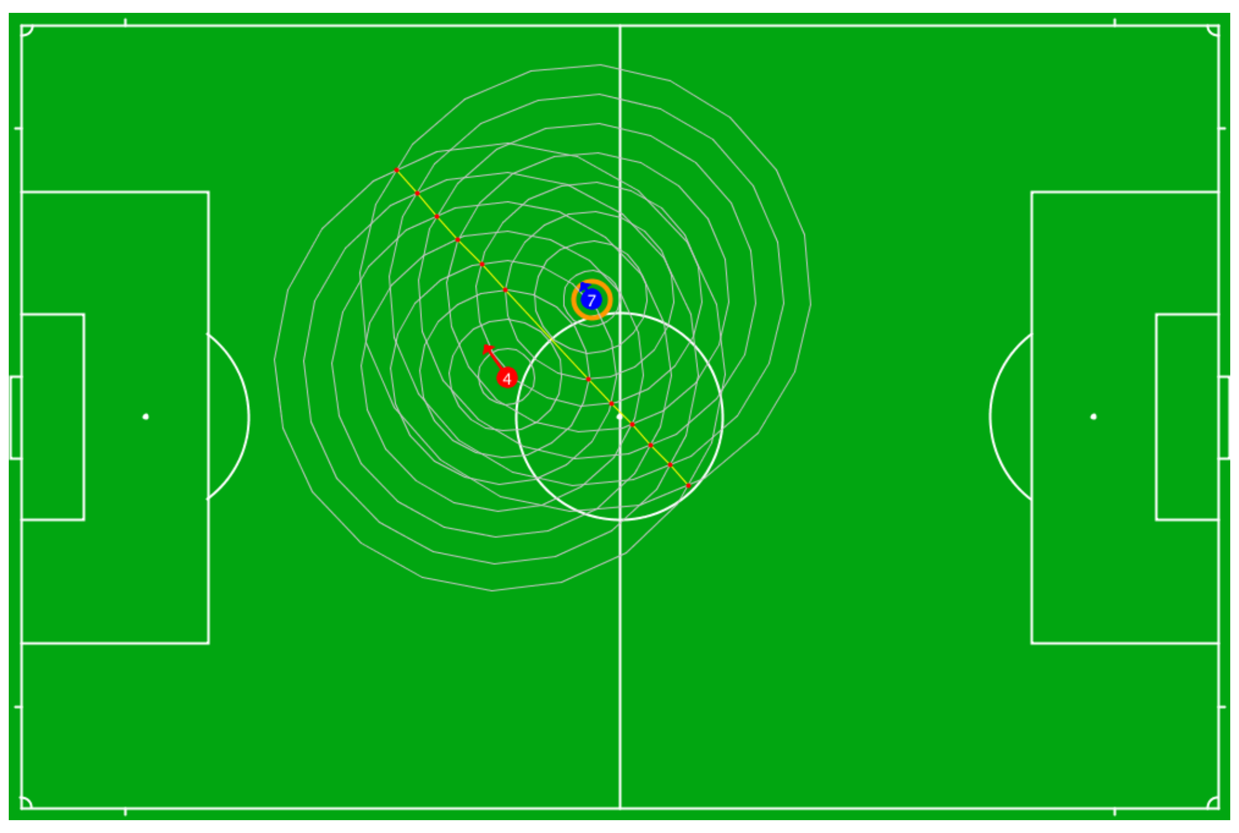
The input data is used to construct a vector of predictor variables for each pass made during the match. In addition, the quality of the passes made in the four matches have been manually labelled by human observers.

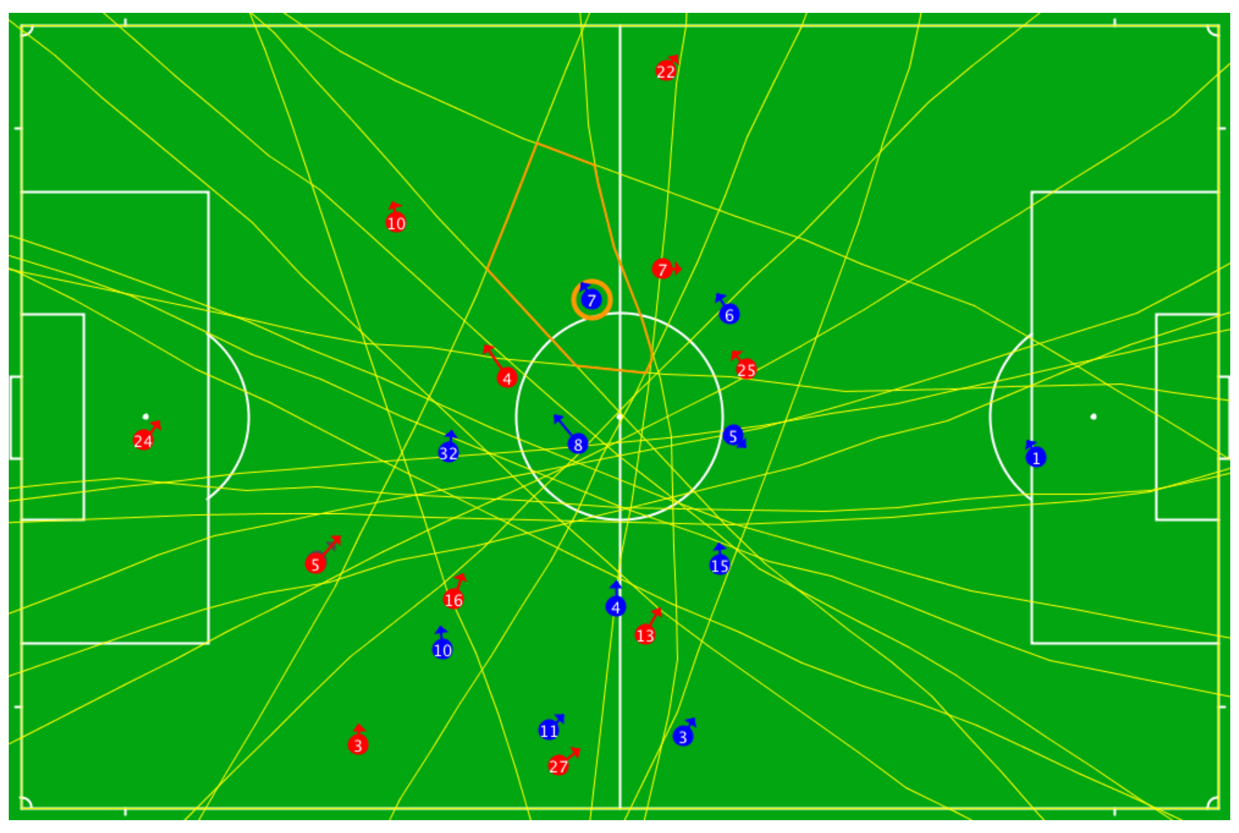
PREDICTOR VARIABLES

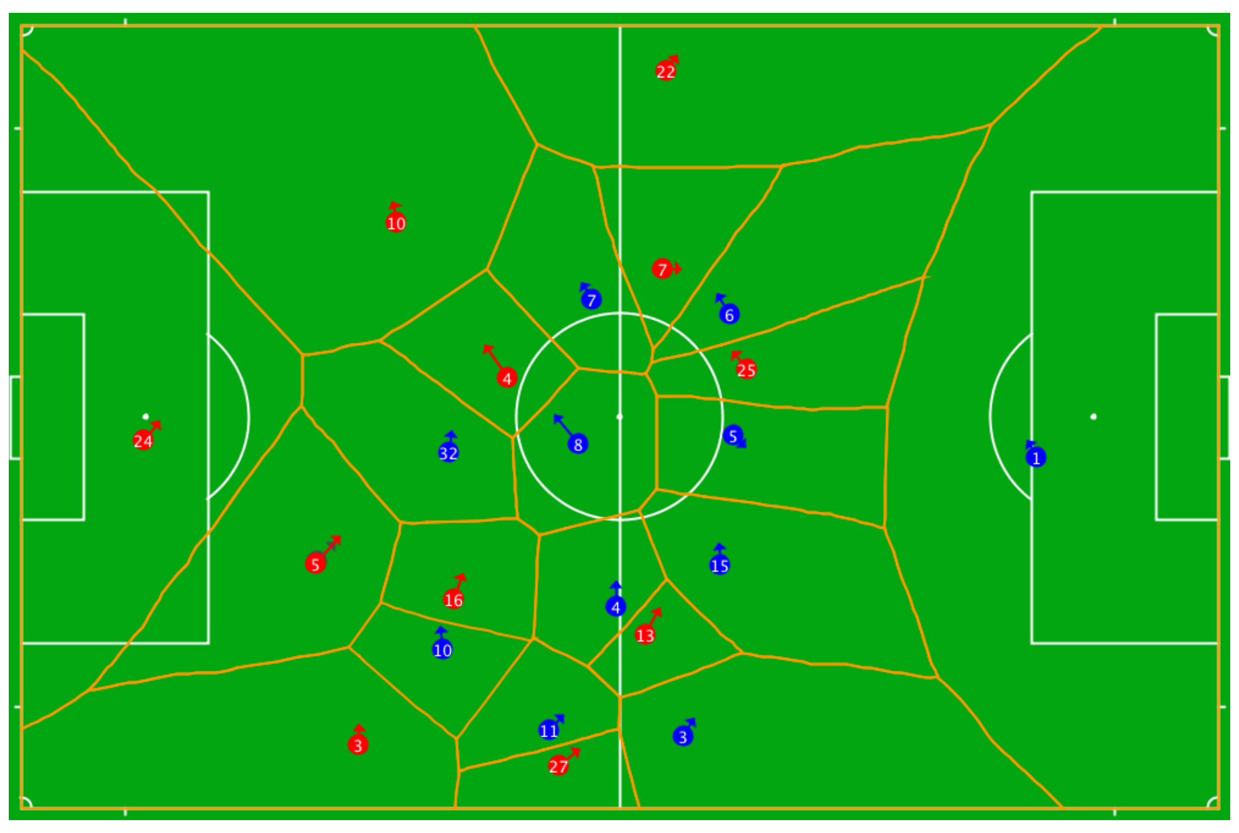
Feature functions are used to compute the predictor variables from the source data:

- **Basic geometric predictor variables** are computed using simple geometric operations such as determining angles between points, measuring Euclidean distances, and calculating velocity of objects over a time interval.
- **Sequential predictor variables** are constructed from the event sequence data.
- **Physiological predictor variables** are predictor variables that incorporate some aspect of the physiological capabilities of the players.
- **Strategic predictor variables** are designed to provide information about the strategic element of the football match.

The physiological and strategic predictor variables required further data structures for their computation. The feature functions used methods from computational geometry with the intention to capture latent structure in the source data.







DOMINANT REGION

Taki and Hasegawa (2000) presents the Dominant Region as a structure to capture information about the tactical position of a football match. The dominant region is a sub-division of the football pitch into regions that are centred around each player, where each region contains all the points that the player could reach before any other player.

We developed an efficient approximation algorithm to construct the dominant region. The points on the pitch that a player can reach are approximated by a motion model that describes boundaries of the areas the player can reach in a given time. The intersection of the reachable boundaries between players determine the constraints of the dominant region (see top figure). By computing the boundaries between a particular player and all others the dominant region for the player is determined (see middle figure). Computing the dominant region for all players produces the subdivision (see bottom figure).

Features can then be constructed from this subdivision, such as the area of receiving player's dominant region, or the net change in the team's dominant region area from a pass.

LEARNING ALGORITHMS

Pass evaluation is a classification task and we evaluate several supervised machine learning algorithms for this purpose. The distribution of example data is skewed, with the majority of examples were clustered towards the middle of the scale. We thus selected learning algorithms designed to handle such class imbalance.

The experiments were run on five classifiers. First, we used multinomial logistic regression (MLR) with three different regularized cost functions: maximum likelihood, arithmetic and quadratic. Second, we used classifiers produced by the Support Vector Machine (SVM) and a RUSBoost algorithms. The intention was to perform the experiments using diverse types of classifiers that explicitly address class imbalance in the data.

EXPERIMENTS AND RESULTS

Overall, we were able to produce a classifier with 86% accuracy on the pass labelling task, see the table, below. Moreover, the inter-rater agreement between the classifier and a human observer was comparable to that between two observers.

The importance of predictor variables based on the dominant region were evaluated by examining the weights vector learned using MLR with L1 regularization. Three of the seven non-zero coefficients produced by this algorithm.

The approach taken appears to produce a useful classifier for this task, and there are several promising areas for further research.

Classifier	Accuracy	Precision	Recall	F-Score
MLR-MLE	0.829	0.666	0.752	0.638
MLR-Arith	0.730	0.580	0.780	0.612
MLR-Quad	0.741	0.581	0.784	0.617
SVM	0.858	0.713	0.734	0.711
RUSBoost	0.756	0.600	0.781	0.646