A Weighted Curve Clustering Approach for Analyzing Pass Rush Routes in American Football

Robert Bajons[1], Kurt Hornik[1]

[1] Institute of Statistics and Mathematics, Vienna University of Business and Economics, Austria

E-mail for correspondence: robert.bajons@wu.ac.at

Abstract: We present a weighted K-means approach for clustering weighted curves, i.e. curves which may be assigned weights at each observation of the curve. The methodology is applied to routes of defending players in American football, where the aim is to automatically detect effective pass rushing routes from specific players or teams. Preliminary results demonstrate that the methodology used is able to cluster pass rushing routes effectively and much better than a classical (unweighted) K-means approach.

Keywords: Sports Analytics; Curve Clustering; Weighted K-Means.

1 Introduction

Recently, as new data formats such as event stream data, play by play data and specifically tracking data have been developed, the problem of clustering curves has found attention in sports modelling. In (team) sports, such as American football or European Football (Soccer), players naturally move on the pitch in specific trajectories. Since usually the paths of players on the pitch follow certain criteria defined by the players position as well as the tactics of the team, interesting analyses can be derived from studying common pattern in these movements. Miller and Bornn (2017) for example studied player trajectories in Basketball by clustering possession into groups of similar offensive structure. Chu et. al. (2020) used similar techniques to cluster routes of wide receivers in football and derive a database of predefined routes.

The main motivation of this work is distinct from previous approaches and is based on the idea that in football and soccer, routes or possessions (or in

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

general trajectories of players/events) can be assigned weights. Instead of considering routes or possessions only as observed (x, y)-pairs of trajectories on the field, they can rather be viewed as a sequence of triplets (w, x, y). An example from football are pass rushing routes from defensive players. It is possible to assign to each observation of the trajectory a pressure probability, which would serve as a weight for each (x, y)-pair. Then, it makes sense to find structure in the weighted trajectories instead of the original curves.

In this paper, we present a weighted K-means approach to cluster the (weighted) trajectories of pass rushing defenders in American football, i.e. defenders, whose aim is to attack the quarterback and hinder him from throwing a pass. We consider a dataset provided by NextGenStats via the NFL Big Data Bowl 2023 competition on Kaggle, which contains tracking data of every player on every passing play from the first 8 weeks of the 2021 season of the NFL. For each player and play the data contains (x, y)-coordinates of the trajectories until some event (usually when the ball is thrown). We first build a model which assigns probabilities of quarterback pressure at (roughly) every timepoint in order to obtain weighted trajectories for each defensive player.

2 Methodology

Formally, we consider data $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, where each $\boldsymbol{y_i}$ is an $m_i \times 3$ dimensional matrix of weighted trajectories, comprising of a vector of x-coordinates, y-coordinates and a vector of weights w. Since m_i is not fixed but varies for each data point due to fact that some plays take longer than others, it is necessary to unitize the data in order to use a K-means approach. We thus approximate each trajectory by a Bézier curve evaluated at a fixed number M of points. Details about this adjustment are omitted to comply with the predefined page limit of the short paper.

We proceed by briefly describing the clustering methodology. The classical K-means approach tries to find an optimal partition of the n observations $(x_i, \ldots, x_n)$ into K cluster g_i, $i = 1, \ldots, K$, such that the within cluster sum of squares

$$S = \sum_{k=1}^{K} \sum_{i:g_i=k} (x_i - p_k)^2 \tag{1}$$

is minimized. The prototype p_k is given as the cluster mean , $p_k = \frac{1}{N_k} \sum_{i:g_i=k} x_i$. In the case of weighted observations it makes sense to take the weights into account by adjusting equation (1) such that instead the aim is to minimize

$$\sum_{k=1}^{K} \sum_{i:g_i=k} v_i (x_i - p_k)^2. \tag{2}$$

The resulting optimal prototypes for a cluster are found as the weighted averages $p_k = \frac{\sum_{i:g_i=k} v_i x_i}{\sum_{i:g_i=k} v_i}$.
To adapt the algorithm to the data at hand, each observation $\boldsymbol{y}_i \in \mathbb{R}^{M\times 3}$ is transformed such that an $\tilde{M} = 2M$-dimensional vector
$\boldsymbol{z}_i = (x_{1,i}, \dots, x_{M,i}, y_{1,i}, \dots, y_{M,i})$ and a corresponding weights vector $\boldsymbol{w}_i = (w_{1,i}, \dots, w_{M,i}, w_{1,i}, \dots, w_{M,i})$ of the same dimension is obtained. The problem is then to find clusters and prototypes such that the following expression is minimized:

$$\min_{(p_{jk}),(g_i)} \sum_{k=1}^{K} \sum_{i:g_i=k} \sum_{j=1}^{\tilde{M}} w_{i,j}(z_{i,j} - p_{k,j})^2. \tag{3}$$

In analogy to classical K-means algorithms we implemented an iterative refinement procedure which is initialized by an appropriate starting assignment of clusters and then alternates between finding the optimal prototypes for given cluster assignments and finding the optimal cluster assignment given prototypes, until convergence is achieved, i.e. the change in the function to optimize is below some tolerance. The optimal prototypes for given cluster assignment are given by

$$p_{k,j} = \frac{\sum_{i:g_i=k} w_{i,j} z_{i,j}}{\sum_{i:g_i=k} w_{i,j}}, \tag{4}$$

whereas the optimal cluster assignment given prototypes is found by minimizing

$$\sum_{j=1}^{M} w_{i,j}(z_{i,j} - p_{k,j})^2, \tag{5}$$

over k.

3 Results

The left frame of Figure 1 shows the result of the weighted K-means algorithm described in the previous section for the defensive football players when using 12 clusters. Note that the X-axis is scaled, such that plays are from left to right (from the viewpoint of the offensive team) and the value 0 indicates the line of scrimmage for the play. The aim is to derive clusters of similar routes, where the weights, representing the probability of putting pressure on the quarterback (QB), are taken into account. In essence the idea is to distinguish between routes with high pressure outcome and low pressure outcome. If the algorithm is able to do so automatically it is possible to identify strengths of players as well as teams. It can be observed nicely that there are outside as well as inside route clusters, and it is possible to identify effective routes and less effective routes (as given by the pressure weights) for both categories. The grey cluster in the left part of Figure 1 (Cluster 8) seems odd at first sight but upon examination it is clear that it comprises of coverage routes. This is particularly nice, as when analyzing pressure on the quarterback, such routes are not of importance and/or interest. In theory one could simply omit them for the clustering exercise, however from the data it is not clear how to distinguish them. Although there are role labels for defenders ("Pass Rush" and "Coverage"), often coverage players also attack the quarterback. To emphasize the importance of using a weighted K-means approach as opposed to a classical K-means algorithm, the results from the latter approach are shown in the right frame of Figure 1. From a pure route specific point of view the clusters seem reasonable, we observe pass rush clusters and coverage clusters. However, there are two main issues. First, when analyzing pressure on the QB, coverage routes are not of much interest as is also evident from the weights, which are (almost) 0 for these route clusters (clusters 2,6,7,10,12 in the right frame of Figure 1). Second, judging from the weights of the pass rush route clusters (clusters 1,3,9,11), we are not able distinguish between more threatening and less threatening routes, so the clustering is useless when trying to identify which teams or players are effective at which position.

References

Chu, D., Reyers, M., Thomson, J., and Wu, L. (2020). Route identification in the National Football League: An application of model-based curve clustering using the EM algorithm. *Journal of Quantitative Analysis in Sports,*. 16(2), 121-132.

Miller, A.C., and Bornn, L. (2017). Possession sketches : Mapping NBA strategies In: *Proceedings of the 2017 MIT Sloan Sports Analytics Conference.*

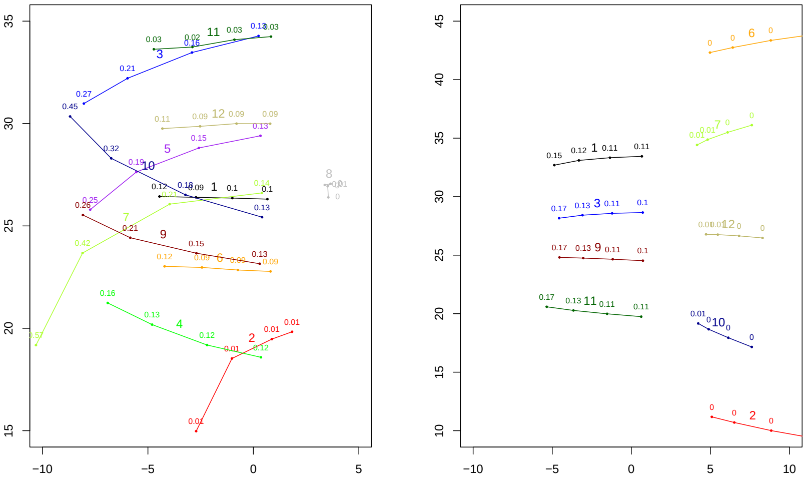

FIGURE 1. Left: 12 cluster as obtained from the weighted K-means algorithm with average weights at each observation point of the trajectory. Right: 12 clusters as obtained from the usual (non-weighted) K-means algorithm (only 9 clusters shown).