SI 670 Final Project Proposal
Mike Verhulst

Determining NHL Player Types Using Clustering Analysis

Introduction

For my final project, I intend to use clustering techniques such as K-Nearest Neighbors to cluster players in the National Hockey League (NHL) based on their statistics from the last few decades. Through clustering, I hope to gain insights on what player archetypes are present in each era and if there is any change in those archetypes over time. The way hockey has been played, especially at the professional level, has changed dramatically over the last several decades and this project aims to quantify that change.

One of my motivations for this project is that conventional hockey wisdom says that there is no room for so-called “goons” at the NHL level anymore. A “goon” is a player whose primary role on a team is to protect his teammates by focusing on dealing out hard hits, acting aggressively, fighting, and being a general nuisance to the opposing team. In most cases, “goons” don’t typically contribute to a team’s offensive production. Most people that follow the NHL will tell you that the league has become so competitive that teams can’t afford to have such one-dimensional players on their rosters anymore.

Clustering has been used in the past to group players by how they play the game[1] but whether or not those play styles have changed over time has not been examined. The goal of this project is to use clustering to extract distinct styles of play of different NHL positions in batches of five seasons. Once clusters have been generated for each of these batches, I will analyze the results to see what (if any) changes in player archetypes have occurred over the last 30 years.

Data

The data for this project will primarily be sourced from hockeyreference.com. This website has team, and player level statistics dating back to the founding of the NHL in 1917. For this particular project I plan on using only the statistics from roughly 1990 to the present. Most hockey experts would define the modern era of hockey as starting around 1990 due to the influx of new teams and players from the former Soviet Union finally being able to join the league around this time.

There will be a fair amount of preprocessing of the data that will need to occur for this project since many of the statistics I’m planning to use are not all stored in the same place. For example, the data for the height and weight of a player is very important in determining their play style but it is stored in separate tables from other statistics such as the goals, assists,and

---

[1] Vincent, Claude B and Eastman, Byron. "Defining the Style of Play in the NHL: An Application of Cluster Analysis" *Journal of Quantitative Analysis in Sports* 5, no. 1 (2009). https://doi.org/10.2202/1559-0410.1133 ; Timothy C. Y. Chan, Justin A. Cho, David C. Novati, (2012) “Quantifying the Contribution of NHL Player Types to Team Performance.” *Interfaces* 42, no. 2 (April 2012): 131-145. https://doi.org/10.1287/inte.1110.0612

penalty minutes. Getting all this data arranged in a way that will be useful for clustering will be a significant portion of this project.

Analysis

The main questions I aim to answer with this project are the following:

1. Does the number of optimal clusters per position change from batch to batch?
    a. i.e., does the number of distinct play styles per position change over time?
2. Do the characteristics of the cluster (i.e., play styles) change from batch to batch?
3. Were “goons” ever prominent enough to have their own cluster? If so, when did that occur?
4. Are there any play styles that only occur once or twice?

Some of the interpretation of the resulting clusters will rely on domain knowledge that can’t necessarily be gleaned from the data alone if the viewer is not familiar with the game of hockey. Certain styles of plays such as the previously mentioned “goons” are more easily defined (such as having a high number of hits and penalty minutes while having relatively few points). However, other clusters may be harder to define without knowing some of the other commonly acknowledged play styles defined by NHL scouts and hockey journalists. These additional play styles will be the basis for my interpretation of the less defined clusters that may occur during this project.

Past Work

Similar work has been conducted in the past and I plan to expand upon this work. Two papers in particular are of note: a 2009 paper by Claude Vincent and Byron Eastman[2] and a 2012 paper by Timothy Chan, Justin Cho, and David Novati.[3] The work by Vincent and Eastman focused on determining which clusters naturally occur and then use available salary data in an effort to compare the differences in salary among these clusters. Similarly, the second paper also strives to find which play styles exist in the NHL with a goal of quantifying how each role contributes to a team's success.

Where I seek to expand upon this work largely relates to the time periods studied by each paper. Vincent and Eastman’s work only considers data through the 2002-2003 season and the work by Chan, Cho, and Novati focuses entirely on seasons after the canceled 2004-2005 season. As noted by Chan, Cho, and Novarti, the rule changes that resulted from the 2004-2005

---

[2] Vincent, Claude B and Eastman, Byron. "Defining the Style of Play in the NHL: An Application of Cluster Analysis" *Journal of Quantitative Analysis in Sports* 5, no. 1 (2009). https://doi.org/10.2202/1559-0410.1133

[3] Timothy C. Y. Chan, Justin A. Cho, David C. Novati, (2012) “Quantifying the Contribution of NHL Player Types to Team Performance.” *Interfaces* 42, no. 2 (April 2012): 131-145. https://doi.org/10.1287/inte.1110.0612

lockout lead to NHL games that were less physical and focused more on offensive output, thus making the comparison of pre-lockout and post-lockout clusters not directly comparable (132).

The claim that pre-lockout and post-lockout clusters are not comparable is what I intend to examine with this project. The two previously cited papers use data that has no overlap as far as NHL seasons are concerned. For my project I plan to bridge this gap by using data from roughly 1990 to the present. By using data that spans the work of both Vincent and Eastman and Chan, Cho, and Novarti, it will give insight to whether or not pre-lockout and post-lockout clusters differ to the point where they are no longer comparable.