Authors:

Oliver Schulte, Yejia Liu, Chao Li, School of Computing Science, Simon Fraser University, Vancouver, Canada

Title: Model-Based Clustering for Identifying Exceptional Players in the NHL Draft  
or  
Tree Models for Identifying Exceptional Players in the NHL Draft

# Introduction

Recruiting strong players to a team is crucial for the team’s success, so predicting the performance of athletes is one of the fundamental problems of sports analytics. We describe a new approach to an important instance of this problem: assessing the prospects for success in the National Hockey League from data available before a player is drafted into the NHL. Two main approaches have been successful: 1) model-based, building a predictive model based on data features [Shuckers], and 2) cohort-based [Weisskopf], deriving performance predictions for a target player from the observed performance of comparable players. This paper develops model-based clustering, which aims to incorporate the strengths and address the limitations of both.

Model-based clustering is a well-established statistical framework [Rafferty]: at the end of the procedure, each individual is assigned to a cluster and *a separate model is built for each cluster.* Compared to a single linear model, the cluster models form an ensemble that increases predictive power. Compared to cohort-based approaches, the clusters are learned from the data based on predictive accuracy, rather than assuming a pre-specified similarity metric. A model ensemble can easily adapt to the variability of different playing conditions, for example by learning a different model for different junior leagues (e.g. OHL vs. WHL) and for different seasons (e.g. 2001 vs. 2007). We utilize a model tree to ensure that the learned player clusters are defined by simple rules that can be easily interpreted by hockey experts. The model tree can be used not only predictively to assess future performance, but also descriptively to highlight exceptional players and their exceptional features (strong or weak points).

# Methods

Since our goal is to support teams in draft decisions, it is very important that the model and its predictions can be explained to and interpreted by hockey experts. As [Florida Panthers] explained, “the numbers are the beginning of the conversation [between scouts and analysts]”. Following [Shuckers], we aim to predict the number *gi* of games that player *i* plays within his first 7 years of NHL play. To ensure that the learned clusters are interpretable, we employ *model tree* techniques rather than optimization. A model tree partitions the feature space according to the values of continuous variables or learned thresholds for continuous variables. Each leaf node in the tree defines a group of players and specifies a separate group regression model. Model trees are flexible and can be employed with any regression model. In this paper we employ logistic regression, where the dependent binary variable is “draftee plays more than 0 games in the NHL”. To each player *i*, the model ensemble assigns a probability *pi* of playing more than 0 games. The zero-count group comprises the players to whom the model assigns probability < 0.5 of playing more than 0 games. Other players are ranked according to *pi*, except that players with *pi* < 0.5 are applied to the zero-count bottom rank. The motivation for filtering out players who are likely not to play any NHL games is that this is a large group (around 50% of a given draft year [Shuckers]). Our hierarchical method therefore offers a new approach to the zero-inflation problem for predicting future games played.

# Results

Figure 1 shows the model tree learned on our data set [describe dataset, post on github].

Our metric of predictive accuracy is the same as that of [Shuckers]: players are ranked by 1) the learned model 2) the number of games *gi*. Then we report the Spearman rank correlation between rankings 1) and 2). (The Kendall rank correlation shows a similarly strong performance.)

Table 2 illustrates the learned clusters and shows the top player in each cluster. To facilitate interpreting the final prediction, we extract for each player their strongest (weakest) feature, which contributes the most to increasing (decreasing) their predictive score.

# Conclusion

Prescriptive vs. descriptive

## Methods

Explain LMT. Show model tree.

[Figure 1. Model tree.] [Figure 2: visualize of 2 clusters in terms of TOI, GP. Maybe pick top 3 players in cluster]

(have to write this)

|  |  |  |
| --- | --- | --- |
| Draft Year | LMT classification accuracy | LMT correlation |
|  |  |  |
|  |  |  |

Mention Shuckers, draft order

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster Definition | Cluster equation | Cluster Size | Top Player | Strongest Point | Weakest Point |
| Show path | Show equation (maybe top three players) |  |  | (e.g. age vs. average age) |  |
|  |  |  |  |  |  |

# Cutting Floor

## How to trade a player: Clustering and Ranking NHL Players

# Sloan Instructions

Abstracts must contain fewer than 500 words, including title and body.

* Abstracts may include up to two tables or figures combined (e.g. 1 figure and 1 table, or 2 tables)
* Each abstract should contain the following sections:
  + Introduction – What question is this research trying to answer? Why is it an important question for the industry?
  + Methods – Description of relevant statistical methods used, including data sources or data collection procedures
  + Results – Description of actual (not promised) results along with relevant statistics
  + Conclusion – The overall takeaway from the study, including how the results will impact the sports industry
* All abstracts must be submitted in one PDF through the [2017 Abstract Submission online submission page](http://www.sloansportsconference.com/activities/research-papers/research-paper-abstract-submission/)

Notes: Sep 18 form OS. Great comments. We’re working on the tables. Should think of a name for the impact. Candidates:

1. Goal Impact

2. Expected Goals Added

3. Goal Probability Added

Each of these is similar to what has been used before.

Average impact per game:

|  |  |
| --- | --- |
| ClusterId | Avg(Impact\_per\_Game) |
| 1 | 2.431 |
| 2 | 2.521 |
| 3 | 0.002 |
| 4 | 0.284 |
| 5 | 2.204 |
| 6 | 2.851 |
| 7 | 2.626 |
| 8 | 2.678 |
| 9 | 0.290 |
| 10 | 2.729 |
| 11 | 3.181 |