

Abstract of Real-Time Optimization using Time Series Data

Barbara Südy

BME, 2014

Internal Supervisor: János Tóth
External Supervisor: Csaba Gáspár

Data mining is used in various domains in real life such as economics, engineering, molecular modeling, finance, etc. Since the early 2000's it has gained more and more popularity in professional sports from baseball to cricket or soccer. In my thesis project I use predictive data mining techniques to build a model which attempts to determine the optimal team structure of an ice hockey team for the upcoming shift. The model gives the coach feedback on the optimal line combination in real time — therefore it provides a novel data-driven approach for in-game decision making in ice hockey.

For building the model I extracted data from the official NHL website where various types of game and player statistics are available for anyone. For the modeling phase I used the game reports of all regular season games of Anaheim Ducks played between October 2009 and April 2012. The reports contain all relevant game related information ordered by time. By using data of all regular season games played in 3 seasons I obtained a large and reliable data set.

The predictions of the model are based on the time series data that arises from the past game events. The time series analysis accounts for the fact that data points taken over time may have an internal structure (such as auto-correlation, trend or seasonal variation) that should be accounted for.

The model consists of three main parts:

1. The first model predicts whether one specific player will or will not play in the next shift.
2. A second model uses the results of the first model as predictors and predicts which players will form the line at the upcoming face-off.
3. The third part, based on the results of the second model gives an alternative line which has at least the same or higher probability for scoring in the upcoming shift.

Structure of the Thesis

Chapter 1 gives a short introduction on the purpose of the study, the basic ice hockey rules, the basic concept of data mining and machine learning and the tools that were used in the project. Chapter 2 describes the data set and the data manipulation methods that were used to transform the data in the desired format. The mathematical background of the classification algorithms is covered in Chapter 3 as well as the model improvement methods and evaluation metrics that were used in the project.

The mathematical description of the model is introduced in Chapter 4. The model consists of three main parts and each part is described in one of the three sections. Starting with a baseline model I introduced the development of the three models, gave a mathematical formalization of the target and predictor variables as well as the output of the model and evaluated the results.

In Section 4.1 I wrote about the first model which applies a binary classification to predict whether one specific player will or will not play in the next shift. The classification reports prove that the final model performs significantly better than the baseline model: for most players, the area under the ROC curve of the binary classification is over 0.8 on the test data, which exceeded my original expectations.

The second part is covered in Section 4.2. It uses the results of the first model as predictors and predicts which players will form the line at the upcoming face-off. This model is driven by a multinomial classification algorithm and assigns a player number to each position. The results show that the classification error rate of the final model is 43%. Although this is not

an extremely good result, it proves that the model definitely functions better than random guessing, and by finding the right way it could be improved. At the end of Section 4.2 I gather some ideas that could potentially improve the model performance.

The third model is described in Section 4.3. Based on the results of the second model, it generates all possible line combinations and returns the one which has maximal probability for scoring in the next shift. Therefore it optimises the team for the upcoming shift. The quality of the binary classification behind the third model is quite poor, the area under the ROC curve is only 0.62 on the test set and the sensitivity is only slightly better than 0.3. I am currently working on finding the way to improve the model by using better predictors and a new target variable in the model. Although the new model has not been properly tested yet, it already seems to show a better performance than the one covered in this paper.

The overall model performance could only be assessed by testing it in real life. Without the tests it is hard to guarantee that the alternative lines would indeed raise the number of goals in long term. However the results of the model parts suggest that – after further development – the model is potentially usable and it could provide the coach a scientific approach in optimizing the line structure in real-time.