NBA Oracle

Matthew Beckler, Hongfei Wang
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
{mbeckler, hongfei}@cmu.edu

Michael Papamichael
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
mpapamic@cs.cmu.edu

1   Introduction

In this project we will apply machine learning techniques to an existing dataset of basketball statistics to: 1) Predict the outcome of a game, given the two participating teams, and 2) Identify outstanding players based on season and career statistics.

2   Data Set

We will be using National Basketball Association (NBA) and American Basketball Association (ABA) statistical data, maintained by *www.databaseBasketball.com*. This dataset contains player and team statistics from all NBA and ABA games through the history of these two leagues. If necessary, we will apply transformations to the dataset to satsify the needs of the specific query under study.

3   Project Idea

The business of professional sports is a multi-billion dollar industry, with the NBA being one of its core constituents, with over $3.5 billion dollars of revenue in the 2007-2008 season. With many teams desperate for a winning season, the importance of well informed decisions regarding player aquisitions [1], based on data analysis, is critical [2]. Betting on sporting events, which is another multi-billion dollar industry on its own, similarly depends on accurate data analysis for predicting game outcomes.

For this project we will apply machine learning techniques to: 1) Predict the outcome of a game, and 2) identfity the best and worst players. To address the first problem, we plan to use binary classification techniques – as seen in class – on data regarding the two participating teams. For the second problem, we will apply outlier detection techniques to individual player data to pinpoint both the outstanding and worst performing players. If necessary, we will apply transformations to the raw player data to reduce the input data dimensionality and increase the accuracy and runtime performance of our algorithms.

We will be using MATLAB to develop our ML algorithms. We will also consider using third-party statistical packages, such as Excel or SPSS, to validate our results.

References

[1] Colet, E. and Parker, J. *Advanced Scout: Data mining and knowledge discovery in NBA data.* Data Mining and Knowledge Discovery, Vol. 1, Num. 1, 1997, pp 121 – 125.

[2] McMurray, S. (1995). *Basketball's new high-tech guru.* U.S. News and World Report, December 11, 1995, pp 79 – 80.