Basketball Analytics Pipeline---From Raw Video to Dynamic Visualization

Lukengu Tshiteya, Wenge Xie, Joe Zuo
Project Manager: Heather Mathews
Faculty Lead: Alexander Volfovsky

Introduction

Currently, data science is widely introduced in both NBA and NCAA as a new way to analyze tactics. Our project aims to analyze ball-holder's movements during an NCAA basketball game.

The dataset comes from SportVu (former NBA video tracking technology provider), including all of the game frames of 2014-2015 Duke Men's Basketball games (24 games in total). Each row in the dataset includes the absolute player and ball position information (in x-y coordinates), game clock, shot clock, and current event label. There are 7 different labels: dribble, pass, shot, touch, free throw, turnover, and rebound.

Objectives

Our primary goal of this project is to establish a well-rounded prediction model of basketball player movements. To be specific, our task is divided into 4 parts: Data cleaning and visualization of SportVu data / Apply different classification methods for prediction / Use Sequential Neural Network and Recurrent Neural Network for the prediction model and evaluate their performances / Build an R Shiny App for interactive prediction

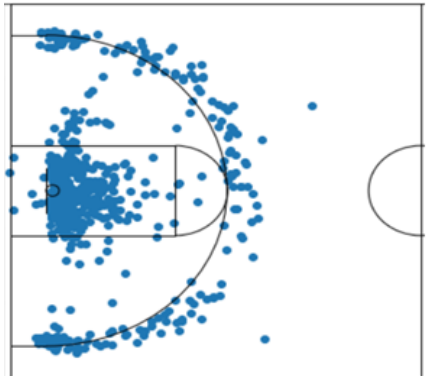


Fig.1 Distribution of Made Shots by Duke players in 2014-2015 season

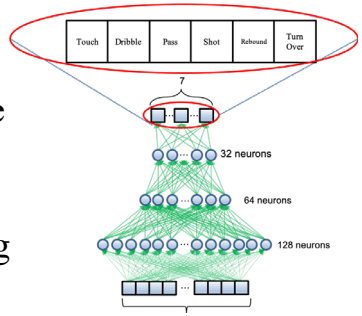
Data Cleaning

- Normalization
  - Organize data into a related table; Also eliminate redundancy and increases the integrity which improves performance of the query
- Class merging
  - Merge 12 classes (in the origin dataset) into 6 classes to make different classes more distinguishable
- SMOTE (Synthetic Minority Over-sampling Technique)
  - An approach to over sample the minority classes (pass, touch,) by creating “synthetic” examples.

Model Building

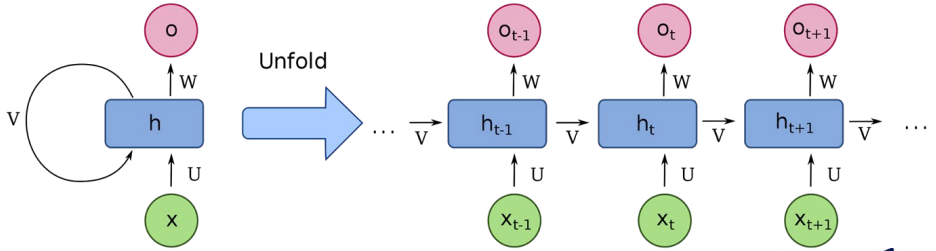
- 4-layer sequential Neural Network

The 49-sized input vector (including the action and position information of the players) will go through each layer in the neural network by inner production. All the parameters in this model will be adjusted well by back propagation during training.



- 4-layer Recurrent Neural Network

Connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs



Duke UNIVERSITY

R Shiny

- Creates and visualizes an interaction database of the event descriptions from the predictive model
- Model is used to identify trends in the 14-15 season and display the constant flux of basketball
- Doubles as tool to predict the offensive identity and makeup of the 14-15 team at certain points and throughout the season

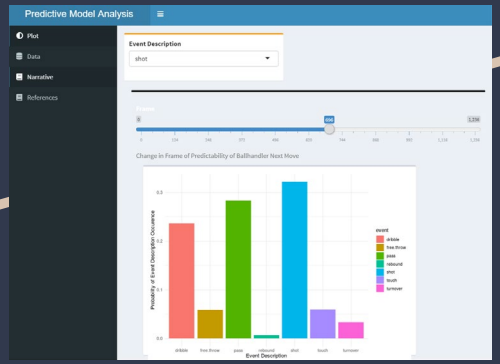


Fig. Frame Screenshot

Conclusion

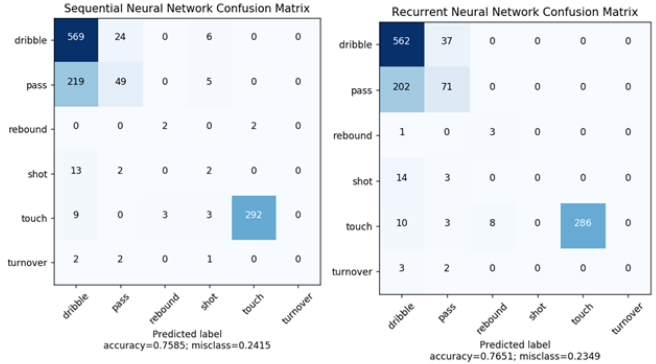


Fig.7 Confusion Matrices of both of our models (The numbers on the diagonal are the numbers of the correct predictions of the corresponding class)

1.Imbalanced data can make the prediction result very biased. In our model, more than 40% of the actions made by the ball holder are dribble. So, our model learned has the trend to predict everything to be dribble. Using normalization and SMOTE can decrease the influence brought by the imbalanced data.

2.RNN has better performance on sequential data. It improved the testing accuracy of our model from 75.85% to 76.51%. More importantly, RNN improved our model's performance on imbalanced dataset. It decreased the error rate of "pass" (the most imbalanced class in our dataset) from 77% to 71%.

3.The reason RNN did a better performance on the minority classes (pass) is because it sacrificed some of the prediction accuracy on the majority class (dribble). In the confusion matrix, the correct labelled "dribble" decreased from 569 to 562 while the correct labelled "pass" increased from 49 to 71.

Duke UNIVERSITY