INFSCI 2915 Machine Learning: Project Proposal

Title: Study on NCAA Basketball Tournaments

Team Members:

Chen Song (chs222@pitt.edu)
Yuxuan Li (yul154@pitt.edu)
Junjia Guo (jug44@pitt.edu)

Description of the system/problem:

This problem and related data comes from a Kaggle competition "March Machine Learning Mania 2017", which aimed to predict the result of 2017 NCAA basketball tournament with historical data of NCAA basketball games since 1985. Since 2017 NCAA basketball games have already finished, we also have all data about it. Therefore we will rely on historical data from 1985 to 2016 to build prediction models, and use the models to "predict" 2017 game results and evaluate the performance of our models.

Mention the type and source of data you will use:

There are several tables in this dataset, including:

1. Teams: team names and corresponding 4-digit id numbers
2. Seasons: season-year, start date of the season and regions of each season
3. RegularSeasonCompactResults (1985-2015): season, number of date from the start date, winning team and its score, losing team and it score, number of overtime periods, and winning team location (home, visiting or neutral)
4. RegularSeasonDetailedResults (2003-2016): more technical details of each game
5. TourneyCompactResults: similar with RegularSeasonCompactResults
6. TourneyDetailedResults: similar with RegularSeasonDetailedResults
7. TourneySeeds: season, seed and team
8. TourneySlots: season, slot, strongseed and weakseed

Explain how machine learning will be used to solve your problem, and your overall approach

Since the dataset is complex, we will need more effort for data exploration and manipulation, feature selection and feature engineering if it is necessary. For example, if we want to use historical winning rate of the teams as a feature, since the number will change game-by-game, we will need to re-compute the winning rate for each game. The features that possibly been included in our models including seed information, time/round in the season, average winning rate and winning score of each team, etc.

On the other side there will be a lot of potentials to build different models from this dataset. Since this is a classification problem (team1 could win or lose), we could use logistical regression, knn, naive bays, SVM and assemble methods to build different models.

Include the main responsibilities of each team member in the project

Possible workflow will include:

- data exploration and visualization
- data preparation
- baseline model construction
- try to enhance baseline model in different ways
- evaluation and discussion

We have not yet decided main responsibilities of each team member since the general situation of the dataset is still unclear. Junjia and Yuxuan will begin will data exploration and visualization, and Chen will begin with data manipulation.