Project Proposal

DSC 478

Team Members:

Brandon Polk

Warren Kress

Dennis Kim

Project Type: Clustering Analysis

Topic: NFL Team Statistics 2010-Present

Datasets:

https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data#spreadspoke\_scores.csv

https://www.pro-football-reference.com

Overview:

The datasets that we gathered contain a wide range of NFL team statistics per season per year including offense, defense, special teams, etc. We will build a data set from this information and apply clustering techniques to see if we can create distinct interpretable classes, for example, passing oriented teams, rushing oriented teams, defensive teams. We will then analyze and report characteristics defining these new classes and report how they may be associates with other variables such as win/loss percentage, playoff appearances, etc.

Analysis Approach:

DM Task 1: Constructing new data set from original data sources

* We are using multiple datasets, each containing variables that may be relevant for individual teams. We will select features from these data sets that we will use to construct a new complete dataset for use in this project.

DM Task 2: Pre-processing

* While some feature selection will have been done through the construction of the dataset, it is likely we will still need to apply some feature reduction/selection techniques before running it through any algorithm. Some ideas that we are brainstorming include the use of PCA on different subsets of the data.
* Sub DM Tasks: Data cleaning, data visualization, feature selection, feature reduction, PCA

DM Task 3: Clustering and Development of New Class Labels

* The data set is largely containing seasonal statistics for individual teams with no useful explicit class labels. Through clustering analysis, we hope to be able to distinguish clear, interpretable classes that will be useful for further exploration.

DM Task 4: Use of New Classes (Classification/Regression?)

* We can cross reference the newly created classes with win/loss statistics such as win loss percentage, playoff appearance, and playoff wins, and explore distributions of these variables across the classes. We may be able to develop a classifier that labels a team based on its given stats and then, based on its predicted class label, predict the likely outcome of the season for that team (playoff appearance, playoff wins, etc.)

Data Schema and Size:

* The data will have at least 320 samples (32 teams across the most recent 10 seasons), however this may be expanded if we use individual game statistics.
* The number of features is to be determined. We may use PCA on related subsets of the data and then recombine these new subsets with the others to construct the final dataset

Plan for Evaluation – Analysis of Results/ Discussion:

* Method to evaluate the clustering quality to be determined.
* Method to evaluate classification likely to be k-fold cross validation

Plan Work Distribution:

* Specifically, to be determined. The bulk of this project will likely be construction of the final data set and pre-processing which we will all be involved with.