COMP721

NBA Game Outcome Prediction and Player Outlier Identification By Use of Machine Learning

Liresh Kaulasar 217079186

Leenane Makurumure 217076701

Nkosingiphile Mkhize 216018127

Jaryn Arjoon 215076277

Ndabenhle Ndabana 216039431

Abstract

The NBA is a worldwide known basketball league, and in a single game there is a great amount of statistical data generated about individual players and the teams, this data is also collected over an entire season. In this project we use different machine learning approaches to make use of this generated data in order to identify outlying players and to attempt to predict the outcome of a game between two teams, for player outlier detection using PIR and PCA we were able to easily identify outstanding players, for game outcome prediction we used an artificial neural network and achieved an MSE of approximately 0.25%, which shows good forecasts against actual outcomes.

Introduction

The NBA (National Basketball Association) is a men's professional basketball league based in America, it has grown in popularity and influence since it was founded in 1946 [1]. As with any professional sport people become interested in finding out which players are the best or which players outperform their peers, people also become interested in knowing which team is most likely to win the game before it's even started, more especially those who bet on professional sports games. In a single basketball game, there is a lot of statistical data generated about each player and the teams playing, this data is also generated and compiled over a season. We now know that statistical data on players and teams exist and have two questions, "Who are the outstanding players?" and "Can we predict the outcome of a game?", in this project we attempt to solve these two problems by making use of the previously gathered data and applying various machine learning techniques to perform outlier detection on the players and predict the outcome of a game between two teams.

Related Work

Attempting to find out which players are outliers or trying to predict the outcome of a game are not new concepts and there has been work done previously on these subjects. In the paper "Prediction of NBA games based on Machine Learning Methods", a feature vector was created using win-loss percentages, point differentials and prediction based on previous games outcomes between two teams, linear regression was then used with these features being used as the inputs and achieved a performance of 67.89%, they also made use of multilayer perceptron achieving 68.44% and maximum likelihood achieving 66.81%, there is also mention in the results, that experts prediction rate is approximately 70% [2]. "NBA Oracle" is another previous work that makes use of machine learning for game outcome prediction and identifying outlying players, as well as inferring the optimal player positions, they use linear regression, logistic regression, SVM's and ANN's for game outcome prediction and achieved up to 73% accuracy, for player outlier identification they generated two new features, namely Approximate Value and Efficiency, using these features a scatter plot was created to easily identify outlying players [3]. "Predicting NBA Game Outcomes" also makes use of SVM, linear regression and neural network regression for game outcome predictions and achieved average accuracy of 62.07%, 63.75% and 64.95% respectively [4]. "Performance Index Rating" or PIR is a statistical formula for basketball created in 1991, it is used to determine the leagues weekly MVP by EuroLeague and is similar to the NBA's Efficiency statistic [5]. "Player Efficiency Rating" or PER was created by John Hollinger, a columnist for ESPN.com, and is a per-minute rating of players, by taking the total of all positive accomplishments and subtracting all negative accomplishments from that which results in the per-minute rating of player performance [6].

Methods and Techniques

Outlier Detection

In basketball history, there is a lot of statistical data generated about each player and team. Namely: player regular season stats, player regular-season career totals, player playoff stats, player playoff career totals, player all-star game stats, team regular season stats, complete draft history, NBA coaching records by season, NBA career coaching records. We used the regular season career totals and player playoff career totals to discover all-time outlying outstanding players. In both these datasets, each player had the following attributes: games played, minutes, points, offensive rebounds, defensive rebounds, rebounds, assists, steals, blocks, turnover, pf, field goals attempted, field goals made, free throws attempted, free throws made, tpa, tpm.

Dimensionality Reduction

The attributes separately do not represent meaningful information that allows us to compare players, so we perform dimensionality reduction. This allows to transform the given dataset from a high-dimensional space to a more meaningful low dimensional space. By reducing the number of features, we make the modelling task less challenging and visualize each players' properties. We use Principal Component Analysis and The Player Index Rating for this task.

Principal Component analysis

Principal Component Analysis (PCA) is an unsupervised linear transformation technique that compresses data but maintains the most relevant information. It does this based on the correlation between features. It maps the direction of maximum variance in high dimensional data into a lower dimension. [7,8].

PCA Steps:

1. Standardize the dataset

PCA is sensitive to data scaling, so we standardize the features before applying them.

2. Compute the co-variance matrix

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

3. Decompose the co-variance matrix into its eigenvector and eigenvalues.
4. Select k eigenvectors that correspond to k largest eigenvalues
5. Construct a projection matrix w using the k eigenvalues
6. Transform the d-dimensional input dataset using the projection matrix w to obtain a new dimensional feature space.

In this project, we reduced the feature space into a 1-dimensional feature space.

Player Index Rating

The Player Index Rating was once used to determine the regular-season MVP in the Spanish ACB league. We create a new feature by extracting it from the given attributes. The Player Index Rating for each player is calculated from the data using the following formula:

PIR = ([pts]+[reb]+[asts]+[stl]+[blk]+[pf]) – (([fga]-[fgm])+([fta]-[ftm])+([turnover]-[dreb]))

Though PIR does not consider the weighting system to determine the importance of each stat, PIR was opted for over PER since PER is much harder to calculate and required information that was not present in the dataset.

Hierarchical Clustering in outlier detection

Hierarchical Clustering is a technique that clusters/groups similar objects such that there is a high variance between clusters and low variance within clusters. Using PIR, a new dataset was created containing only the player names and their PIR. This data was converted into a .arff file for use in Weka. In Weka, hierarchical Clustering was used to identify outlying players.

Z-Scores in outlier detection

Z-Scores are the number of standard deviations above and below the mean that each value falls. To detect outliers, we found observations with a Z-Score of K standard deviations above the mean and others with a Z-Score of -K standard deviations below the mean, where K = 3 [9,10,11].

The Z-Score for each observation is given by:

$$z = \frac{X-\mu}{\sigma}$$

Interquartile Range in outlier detection

Interquartile Range (IQR) shows how the data is spread around the median and is given by subtracting the first quartile from the third quartile.

IQR = Q3 - Q1

Outliers are then any values below the lower bound and above the upper bound. Where:

lowerBound = Q1 - (1.5 * IQR)
upperBound = Q3 - (1.5 * IQR)

Game Outcome Prediction

Dataset

To predict the game outcome, we used the team regular season statistics with the features: o_fgm, o_fga, o_ftm, o_fta, o_oreb, o_dreb, o_reb, o_asts, o_pf, o_stl, o_to, o_blk, o_3pm, o_3pa, o_pts, d_fgm, d_fga, d_ftm, d_fta, d_oreb, d_dreb, d_reb, d_asts, d_pf, d_stl, d_to, d_blk, d_3pm, d_3pa, d_pts, won, lost. We used stats from 1976 to 2003 as our training dataset and 2004 as the test set. Stats before 1976 were not included because steals, blocks and turnovers were not official NBA stats until the 70’s. The data is pre-processed to reduce sparsity and eliminate bias.

Model

An artificial neural network was used to predict the game outcomes. The input feature vector is a vector of 31 features namely: o_fgm, o_fga, o_ftm, o_fta, o_oreb, o_dreb, o_reb, o_asts, o_pf, o_stl, o_to, o_blk, o_3pm, o_3pa, o_pts, d_fgm, d_fga, d_ftm, d_fta, d_oreb, d_dreb, d_reb, d_asts, d_pf, d_stl, d_to, d_blk, d_3pm, d_3pa, d_pts, pace. The output is the probability of a team wining. Given two teams, we then use this trained model to predict, and the team with the highest probability is predicted to win.

Model Design

The artificial neural network is a three-layer network. Consisting of an input layer, which receives a vector of 31 features, hidden layer with 7 nodes and a ReLU activation function. The model uses the mean squared error loss function and an adam optimizer.

Results and Discussion

Player Outlier Detection

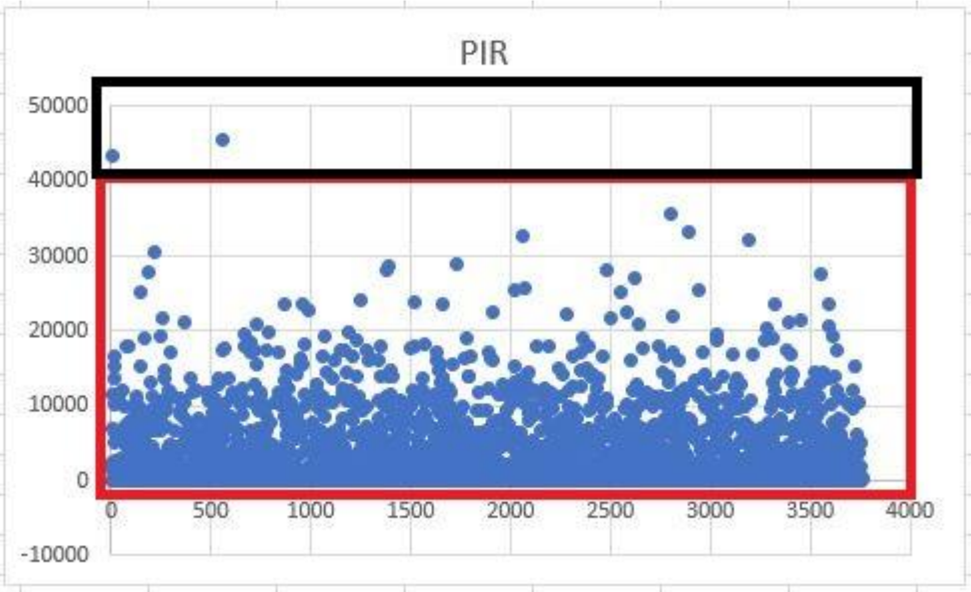

Fig 1. Outlier detection results using PIR

After using PIR for outlier detection it was easy to see from the scatter plot which players were outliers, however when using hierarchical clustering in weka it only identified 2 players as outliers in one cluster (these players were Wilt Chamberlain with the highest PIR and Kareem Abdul – Jabbar with the second highest PIR) and placed all remaining players in another cluster as shown in Fig 1, After using PCA and PIR with z-score, we noticed that both perform the task of identifying outlying players well, and common outliers were detected from both methods like Kareem Abdul – Jabbar and Ray Allen, showing approximately 24 outliers for regular season career totals and 14 for playoff career totals. However, PIR with z-score is more preferred since it takes into account the basketball statistics that PCA does not.

Game Outcome Prediction

For game outcome prediction the model was trained over 30 epochs and we achieved an mse of 0.2538 which shows that the model leans towards better predictions overall, predicting closer to the actual outcomes. However we do not believe that this performs on par or better than previous work done to predict game outcomes.

Conclusion

In this project we used different machine learning approaches to solve the problems of identifying outlying NBA players and predicting the outcome of a game, for the player outlier detection we used various approaches and noticed that there are slight differences in results between them but some common outliers are identified throughout all methods, for game outcome prediction we managed to achieve an mse of approximately 0.25%, which shows that the model does forecast outcomes close to the actual outcomes without overfitting.

References

[1] https://en.wikipedia.org/wiki/National_Basketball_Association

[2] Torres RA. Prediction of NBA games based on Machine Learning Methods. University of Wisconsin, Madison. 2013 Dec.

[3] Beckler M, Wang H, Papamichael M. Nba oracle. Zuletzt besucht am. 2013;17(20082009.9).

[4] http://cs229.stanford.edu/proj2017/final-reports/5231214.pdf

[5] https://en.wikipedia.org/wiki/Performance_Index_Rating

[6] https://www.basketball-reference.com/about/per.html

[7] https://www.youtube.com/watch?v=03Cv8Fc2-tU

[8] https://en.wikipedia.org/wiki/Principal_component_analysis

[9] https://www.youtube.com/watch?v=rzR_cKnkD18

[10] https://www.statisticshowto.com/probability-and-statistics/z-score/

[11]https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561