NBA Play-off & Regular Season Relation Analysis with Artificial Neural Network and Primary Components Analysis

Burak Kaan Bilgehan[*], Cemal Erat[†], Ilayda Beyreli[‡] and Yekta Demirci[§]
[*†]Department of Computer Engineering, [‡§]Department of Electrical & Electronics Engineering
Middle East Technical University
Ankara, Turkey
Email: [*]kaan.bilgehan@ceng.metu.edu.tr, [†]erat.cemal@ceng.metu.edu.t, [‡]ilayda.beyreli@metu.edu.tr, [§]yekta.demirci@metu.edu.tr

***Abstract*—In this paper, we used a predictive learning model for investigatinfg the relation between regular season and play-offs using game data from the National Basketball Association (NBA) and predict the play-off game results in the existance of a relation between these two season phases. During model generation, the correlation between winning probabilities of teams in individual games are taken into consideration while some intangible information of the game setup conditions are integrated by using a scemantic approach.**

I. Introduction

According to TV broadcast statistics , 15 million people in USA watched NBA Finals each year, on average. [1] As a result of combining this statistics with ticket prices and sports betting, the NBA teams and the players hold a high commercial value in the sports industry. When online betting websites are examined, it is seen that there is a high demand on statistical analyses, tips and score predictions on the market. In this paper, we seek the relation between regular season and play-off games and discuss the availability of a model for predicting the play-off tree on which most probable matches among the teams on play-offs and to present the most probable winner of each match with related probabilistic measures. The model has considerable market value in the game betting industry since by changing the features under consideration the model can be generalized for many sports and contests.

There are several points which creates a great challenges for generating a predictive model in such domain. First of all, the very essence of the database includes human factor since when the result of a match is tried to be predicted, it is needed to operate on a player-wise analysis base. Since human behaviour depends on many correlated variables, it is lack of semantically rich representations. Therefore, for such an analysis, imperceptable data like mental-emotional status of the players, can be overlooked easily. However these features play a big role in the performance and they cannot be excluded from the model. Hence, during the model generation, even though, these features will not be directly modelled, by adding some constant, it is aimed to minimize such random error causations.

Another encountered difficulty is that matches are stochastic processes. Future dependent cases like injuries, player ejections may highly affect the final result of a single game as well as the whole season. However, such events are hard to be modelled and predicted beforehand since they do not obey a deterministic mathematical model. In order to obtain a predictive learning model as realistic as possible, such future depending events must also be taken into account.

The huge data pool also presentes a heavy computational load problem and possibility of overfitting. NBA data set offers many features and examples, where some of them are unnecessary for the final prediction. Feature elimination is needed to be done to avoid unnecessary calculations and to reduce complexity. However, it is not easy to foreseen which features have significant impacts for the final result. Therefore, several algorithms are needed to detect features with low impacts. As a result of such elimination, optimization can be done to obtain a faster working algorithm.

II. Related Works

There's no actual "solution" for these type of problems. You can never predict the scores 100%, in fact one of the most popular betting sites in the USA -Vegas- can only predict up to 60-70% of the games, which is a pretty good result. Vegas predictions are often used as a reference point in similar type of projects. Since basketball is a team sport, there are lots of factors we can't define using only numbers. The players can differ in both physical and mental conditions each day. Also the team sinergy can change, the tactics that coaches create might not work every day, injuries may happen. So while deciding which statistics are the most important in the basketball game, we also have to deal with those intangibles. Even though there's no way to indicate those in numeric ways, there are some information we can use. For example, the USA is a huge country, and each NBA team plays 82 games in nearly 160 days. That means the teams will travel a lot during the season. If a team plays after a long trip, it might affect their performance. Another example is back-to-back games (two games in a row). Nearly 14 back-to-back games per season is the average per team in the NBA. If a team is losing many of their back-to-back games we can understand that they might perform better with a better schedule, and in the play-off's there's no back-to-back games. We can increase the number of examples like that. One of the main reasons why the similar project fail most of the time is the intangibles that can't be numerically described. Another reason is, people sometimes use unrelated statistics or titles as features. For example one of the similar projects used MVP (Most Valuable Player of the NBA, only one team will have a player with the MVP accolade), as a feature. Only one team in the 30 teams will have that feature, not to mention only one player in the 450 players will have that. So it will not be a valuable feature for us, in addition to that the MVP is not an incredible player that changes the game everytime, he's just the player that's thought to be the one with the best performance that season. In conclusion we will try to eliminate such useless features, while trying to take care of the intangibles. We will also make use of the advanced stats such as PER, VORP, Win Shares etc. These are the stats that calculated using different stats of players or teams to have more meaningful stat categories than just basic stats such as Rebounds, Assists, Points. A player can score 30 points per game, which is a very high number, but he can be doing it with just 30% shooting which means he's wasting most of the attacks his team has.

III. Methodology

We have decided to work on advanced stats of NBA since they give a more reliable insight about the game performances.The advanced stats are in fact the useful statistics derived from basic statistics like rebound, shooting etc. For example, an important stat True Shooting TS% is basically a weighted average of free throw percentage FT%, field goal percentage FG% and 3-point shooting percentage3P%. Retrieved the data for 30 teams from the online database owned by Sports Reference [2] is collected with *BeautifulSoup* package supported by Python. Regular season data of the 2017-2018 season has been used to start the model since it gives a complete data collection while the data from 2017-2018 season is considered for demostration purposes. Indivisual player statistics are neglected in this since the players may change several times during the season interval and it increases the complexity of the constructed model. Initially, we used 14 advanced stats from both teams (28 stats in total) as the features, and the winner (home:1, away:0) as the label.

During preprocessing phase of the project, we observed that some of the features are complement of each other or can be derived from one another such as Offensive Rating of one team and the Deffensive Rating of the opponent. Also, some of the features are directy linked to the label n such a way that it may deviece the model. Pace and Offensive Rating can be given as an example for this case. Pace is the total number of possesions by the team,and Offensive Rating is the score created by 100 possesions.So if you multiply Pace by Offensive Rating and divide to100, you will get the score which is directly linked to the output label. Hence, for avoiding such cases, some of the features are discarded from the data set manually befor moving forward.

Three paremeters should be set before constructing the model, the number of layers, activation functions and the number of neurons for each layer. The first two is relatively easier to answer. Number of layer in an *artifial neural network* (ANN) is highly related to the number of input features. As expected, a high number of input features requires a large number oflayers. Hence, since related data set has a relative mid-range dimentions, the ANN model constructed to be a two layer model, one input and one output layer. The activation function for the output layer is selected to be *sigmoid* function due to the output properties and the activation function for the input layer is selected to be *tanh* since it put more stress on negative data.

Deciding the number of neuraons is slightly more complex. The number of neurons in the output layer is set to be the avarege of the number of input dimentions and the number of classes. However the input layer may have any number of neurons based on the selected features to be fed to the model. For selecting the optimum nuber of input features that gives the most uselful data, *primary component analysis* (PCA) is applied before training the data. PCA applies the otrhagonal transfortion to the input given in Eqn 2 where X is the input feature matrix and the W is the weight matrix whose components are calculated as in Eqn. 1.

$$w_k = argmax_{||w||=1}||\hat{X}_k w||^2 \tag{1}$$

$$T = XW \tag{2}$$

After finding transform T, the varience ratio for each of the components are examined and the first 15, 10, 8 and 5 components are choosen and used for constructing the model to determine the one wtih the highest accuracy.

During the evaluation phase, the binarycross entropy and the *Adam Optimizer*are used. Adam optimizer is mainly based on stochastic gradient descent and selected due to its straight forward implementation and computational efficiency [3] while binary cross entropy is applicable for binary classification cases.

For the first experiments all the matches are divided into two subsections for each season. Initial 1230 games are used for training and the other 700 mathes are used for test purposes in order to have more stable results while tuning the model parameters. Later on, the averages for each teams are calculated season-wise and fed to the model constructed for that season in order to analyze the model performances on prediction of future games which was the main goal of the project.

IV. Results & Discussions

Starting with previously labeled games for tests during tunning, Fig. 1. illsutrates the training accuracies through number of epochs for different number of selected PCA components to be fed for 2016-2017 season. As expected, when the number of PCA components are decreased, the accuray also decreases since the input feature are not efficiently used. However, decreasing the number of components also decreases the computation time. Hence, we need to select an optimal point of opeation in this trade off. Halving the number of components were giving a train accuracy around 90% which is a considerably good performance while decreasing the computational load significantly. This situation also implies that some of the statistics that are used are more influential than the others such that when the ones with lower influentions are discarded there is not a dramatic change in the model performance as given in Fig. 2.

Average statistics for a team calculated using the regular season results are not a complete indicator of the same team on the same game. Due to this fact, the model is expected to have a lower performance metrics when only average stats are supplied and a prediction is expected. When such a procedure is applied the results in Table I and Table II are observed. Since the games were assumed to be playing or have not been played yet during the prediction, the actual winner field is not added. Being compared with the actual results the final accuracy is very small. For Season 2016-2017 the accuracy for the first round is approximately 0 and for Season 2017-2018 this number becomes approximately 2/8 since the model predicts that both Toronto Raptors and Boston Celtics will be on the next round.

These results support the idea that dynamics in NBA teams are highly affected from non-numeric features. The underlying reasons may be gathered under two parts. First issue that has to be considered is that the home-away feature may be more effected than most people think. When the results of thwo seasons are examined, the model seems to have a tendency towards guessing home teams more successful. Thus, this supports the non-writen belief of the *home-team advantage*. Home teams seems to be 35% more advantageous than their opponents.

Another idea that has been supported by these experiments is that the regular season average and the play-off performance of a team are not highly related. In other words, the game performances of teams seems to have a high varience so that the mean of the performance distribution of teams does not give sufficient information aout indivisual matches. Hence, more stochastic models should be use for evaluating the team performances and games.

V. Conclusion

This study focuses on the team behaviour and performance in ports area and the mathematical modelling of non-numeric features rising from human factor. Hence, more complex models should be used for including semantic data in prediction algorithms in order to fully integrate all the features of the input data. Although PCA is a very useful tool for decreasing the computational load of the machine learning algorithms, it simply seems not to be effective to uncover all the patterns, especially the ones due to intangible effects, in a data set for prediction. Similarly, despite the fact that simple ANN models

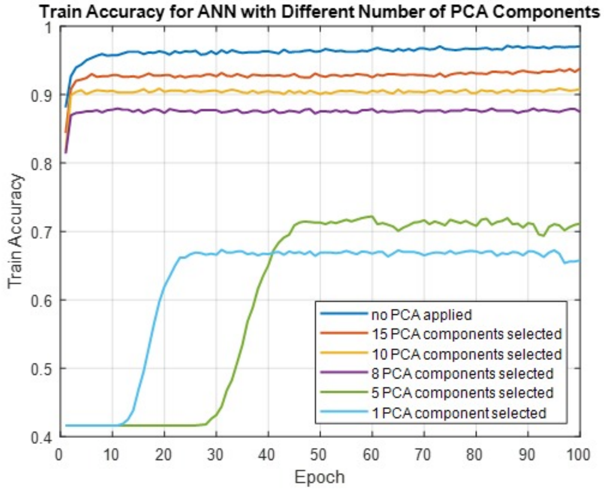

Fig. 1. The train accuracy for the ANN with different number of PCA components

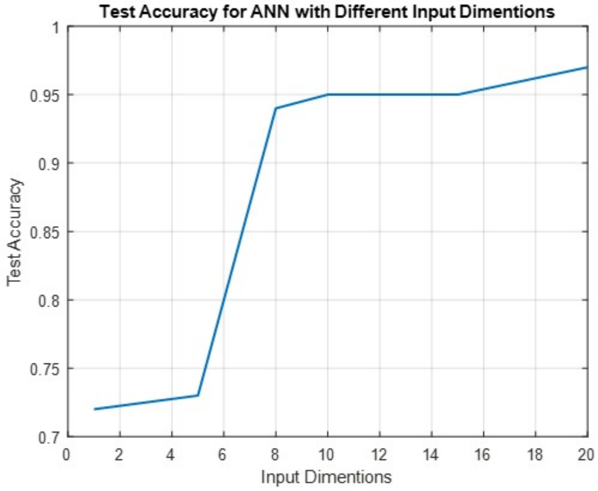

Fig. 2. The test accuracy for the ANN with different input features

TABLE I
Predicted Results for First Round of Play-offs in 2016-2017 Season

Season 2016-2017 Home Team	Away Team	Predicted Winner
Golden State Warriors	Portland Trail Blazers	Home
Portland Trail Blazers	Golden State Warriors	Home
Boston Celtics	Chicago Bulls	Home
Chicago Bulls	Boston Celtics	Home
San Antonio Spurs	Memphis Grizzlies	Home
Memphis Grizzlies	San Antonio Spurs	Home
Cleveland Cavaliers	Indiana Pacers	Home
Indiana Pacers	Cleveland Cavaliers	Home
Houston Rockets	Oklahoma City Thunder	Home
Oklahoma City Thunder	Houston Rockets	Home
Toronto Raptors	Milwaukee Bucks	Away
Milwaukee Bucks	Toronto Raptors	Home
Los Angleles Clippers	Utah Jazz	Home
Utah Jazz	Los Angleles Clippers	Home
Washington Wizards	Atlanta Hawks	Home
Atlanta Hawks	Washington Wizards	Home

TABLE II
PREDICTED RESULTS FOR FIRST ROUND OF PLAY-OFFS IN 2016-2017 SEASON

Season 2017-2018		
Home Team	Away Team	Predicted Winner
Houston Rockets	Minnesota Timberwolves	Away
Minnesota Timberwolves	Houston Rockets	Away
Toronto Raptors	Washington Wizards	Home
Washington Wizards	Toronto Raptors	Away
Goldan State Warriors	San Antonio Spurs	Away
San Antonio Spurs	Goldan State Warriors	Away
Boston Celtics	Milwaukee Bucks	Home
Milwaukee Bucks	Boston Celtics	Away
Philadelphia 76ers	Miami Heat	Away
Miami Heat	Philadelphia 76ers	Home
Cleveland Cavaliers	Indiana Pacers	Home
Indiana Pacers	Cleveland Cavaliers	Home
Portland Trail Blazers	New Orleans Pelicans	Home
New Orleans Pelicans	Portland Trail Blazers	Away
Oklahoma City Thunder	Utah Jazz	Away
Utah Jazz	Oklahoma City Thunder	Away

can find relations in game by game statistics, they are also not sufficient enough to display incomprehensible data and give a single output by processing data of a long period of time.

REFERENCES

[1] NBA Finals average US TV viewership 2002-2017 — Statistic. (n.d.) [online] Available at: https://www.statista.com/statistics/240377/nba-finals-tv-viewership-in-the-united-states/ [Accessed 31 March 2018]

[2] Basketball-Reference.com. (2018). Basketball Statistics and History — Basketball-Reference.com. [online] Available at:https://www.basketball-reference.com/ [Accessed 10 May 2018]

[3] D. Kingma and J. Ba, ”ADAM: A Method for Stochastic Optimization”, in ICLR, 2015.