A novel way to Soccer Match Prediction

Jongho Shin
Email: jongho@stanford.edu

Robert Gasparyan
Email: robertga@stanford.edu

***Abstract*—Our hypothesis is that the video game industry, in the attempt to simulate a realistic experience, has inadvertently collected very accurate data which can be used to solve problems in the real world. In this paper we describe a novel approach to soccer match prediction that makes use of only virtual data collected from a video game(FIFA 2015).**

Our results were comparable and in some places better than results achieved by predictors that used real data. We also use the data provided for each player and the players present in the squad, to analyze the team strategy. Based on our analysis, we were able to suggest better strategies for weak teams

I. Introduction

Sports betting has motivated many machine learning and rule based attempts to solve this problem over the years but one of the most common approaches was using some combination of previous match results as a feature set. In this paper we attempt a novel approach to Soccer Match Prediction that doesn't use match statistics. This, as far as we know, is the first attempt to predict soccer match outcomes using only "Virtual Data"(Video game data). Our goal is to build a better or at least comparable predictor than the ones based on "Real Data" which will verify our hypothesis about the validity of video game data.

Video games have long become a multi-billion dollar industry for a reason none other than the fact that they are really "good". And the reason they have become so good is because a lot of resources have been spent on making them as realistic as possible which requires collecting and curating a lot of real world data. Of course this data has been collected for the sole purpose of making the video simulation more realistic, thus more engaging for the user. A representative of the a soccer video game company said , "The information gathered by our network of more than 1,300 scouts around the world, combined with Prozone"s amazing performance data, makes this an invaluable tool for any football club that takes player recruitment seriously". If this statement is true, and the video games are that accurate the question becomes: Can data collected by video games be used for solving other problems in the same medium? In this paper, we try to answer exactly this question.

The medium we looked at is European soccer and the video game considered is "FIFA 2015". Makers of this video game are famous for creating the most realistic soccer player avatars that look and act very much like real ones. In order to provide this level of authenticity, experts have been hired to collect information about the playing style of each player and curate this data so that each player can be accurately represented as a combination of 33+ features. This features include both physical(speed, power, acceleration, etc) and technical skills(dribbling, heading, accuracy, etc) represented as an integer from 1 to 100. These features are updated constantly as the form of players change over the season. It is interesting to point out that this data is already being used to solve other problems - Scouts in English Premiere League use video games to look for young, promising players with specific skills.[1]. And not too long ago a video game player was appointed as a general manager for a real soccer team[2].

The remainder of this paper is structured as follows. In section 2 & 3 we describe the data sources for this project and preprocessing. Section 4 describes our model selection for match prediction and results. In section 5 we used unsupervised learning techniques to identify the strategy used by the teams based on their squads.

A. Previous Work

Sports prediction is obviously a very hot topic and has always packed the interest of sports fans. Also, in countries where sports betting is legal, sports prediction is as critical as predicting stock prices. But despite this big interest in sports forecasting, it was hard to find serious published work on the subject. This may be due to the fact that "Sports Prediction" is a bit too far from academic interest. We were able to find a few papers from published in Economics journal. One paper [3] tried the multi-layer perceptron for sports prediction, and explains how hard it is predicting sports games. [4] investigated three other prediction methods and compared their accuracy. Their result shows that market prediction and betting odds provide much better forecasts than tipsters. [2] anaylized a sports prediction market of the FIFA World Cup 2006, and compared their prediction accuracy based on history to a random draw. The paper shows that history based prediction is better than a random prediction.

Interestingly, it is more easy to find many sprouts prediction papers from other machine learning classes. Students tried to predict their favorite sports game using machine learning. Many different prediction algorithms have been used but they all had one thing in common: they tried to predict match outcomes based on previous match results. We didn't find any attempts at strategy analysis or any instance of unsupervised learning. This is mainly due to the limitation of their feature sets (previous game history), they can only have small number of features. And small number of features hide many factors of games. Thus they could use only a small part of Machine Learning.

[1] http://www.theguardian.com/football/2014/aug/11/football-manager-computer-game-premier-league-clubs-buy-players

[2] http://www.dailymail.co.uk/sport/football/article-2340324/Football-Manager-Vugar-Huseynzade-got-FC-Baku-job.html

TABLE I: Game data feature list

Attacking	Skill	Movement	Power	Mentality	Defending	Goalkeeping
Crossing	Dribbling	Acceleration	Shot Power	Aggression	Marking	GK Diving
Finishing	Curve	Sprint Speed	Jumping	Interceptions	Standing Tackle	GK Handling
Heading Accuracy	Free Kick Accuracy	Agility	Stamina	Positioning	Sliding Tackle	GK Kicking
Short Passing	Long Passing	Reactions	Strength	Vision		GK Positioning
Volleys	Ball Control	Balance	Long Shots	Penalties		GK Reflexes

II. Dataset

In this project, we looked at games played in Spanish 'La Liga'(Primera Liga). There are 20 teams in the league who play each other in a round-robin fashion twice(A plays B twice, once at home and once away) which adds up to 380 fixtures a year. For this project, we aquired collection of data from three different sources:

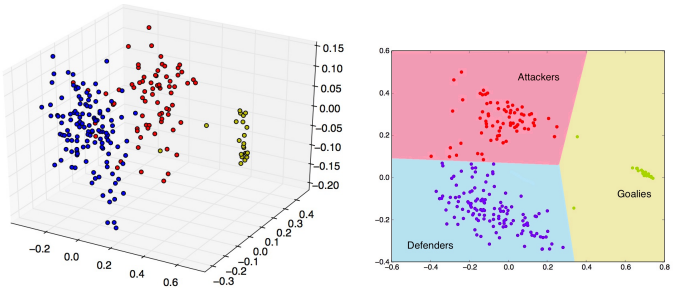
1) Fixture results
   This represents the results of the played matches of 'La Liga'. It corresponds to the output variable in the prediction algorithm. The scores where reduced to a binary value of three categories: home win/draw/away win. Also we collected line ups for each match. We focused on line ups not teams, because teams tend to use different line ups depending on strategy.
   They were gathered from *http://www.goal.com* by scripting web-crawlers.
2) Real Data
   This data represents statistics about teams performance from match history: goals, shots on target, yellow cards, etc. We have looked at many examples of sports prediction techniques and they all use data similar to this. This data was used in this project to create a create a baseline prediction algorithm that our new technique(using virtual data) can be compared against.
3) Virtual Data
   This data was collected from *http://sofifa.com*, and represents 33 features for each players set by experts. Each feature corresponds to a physical or technical skill of player and is an integer value ranging (1, 100). These are the features used by the video game to simulate the actions of each player and in this paper we try to verify how legitimate this data is, by training the predictor with it and comparing the results to the predictor that uses "Real Data". The full set of "Virtual Features" can be found in Table I.

Notice that Real Data and Virtual Data represent features for the same output variables in Fixture results.

III. Features and Preprocessing

A. PCA

Sanity Check with PCA - "Virtual Data" was collected from sofifa.com using web-crawlers, and represents a 33 feature vecotor for each player in the league(total of 280 players). Our hypothesis was that this data was sufficient to build a legitimate predictor for real match results but before we would start creating models and training we decided to do a basic "sanity check". Being familiar with the league and some of the players we wanted to see if the scores given to the players are consistent and are able to distinguish different type of players from each other. In order to make the manual comparison of 280 players in 33 dimensional feature space more feasible we applied PCA. This was a successful and a rather interesting use of PCA, because by reducing the dimensions of the data to 3D(a visually observable space) we could actually look at our data and do a basic sanity check. First thing we noticed after plotting the results of PCA(Figure 1) was a small set of points separated from the rest. We checked and these points correspond to all the goalkeepers in the league which is very intuitive because soccer goalkeepers require a very different from set of skills from field players. Examining the results in Figure 1 further we noticed that the top two players of the league ended up next to each other on the 3D plot. In fact, in the reduced dimensional space, it was easy to see that the players that are close to each other are similar and the ones that are far away are not. We then ran k-means clustering on the data and got three distinct groups (Defensive Players, Offensive Players, Goalkeepers).



(a) PCA in 3D (b) PCA in 2D

Fig. 1: Each point represents a soccer player. The small separated chuck of players are the goalkeepers.

B. Real Data

"Real Data"collected from *http://www.football-data.co.uk* represetns a set of 380 matches played between 20 teams. Each match entry has 24 different statistics(home red cards, away red cards, home team shots on target, etc). This data is used to create a baseline predictor that our "virtual" predictor can be compared against. The aggregation function basically averages teams performance statistics over all the other matches.

$$Team\_Stat = (Home\ Team\ Shots, Away\ Team\ Shots, \ldots, Away\ Team\ Red\ Cards)$$

$$\textit{Team\_Real\_Features(Team)}_i = \frac{1}{m}\sum_{j=1}^{m} \textit{Team\_Stats(Team)}_i$$

Team_Stat represents 12 statistics for a team for a given match. Team_Real_Features is computed by averaging this stats for all the other matches it has played in.

C. Virtual Data

"Virtual Data"represents individual soccer players but we want to predict an outcome of a match played between two squads(a squad consists of 11 players). We define an aggregation function that represents a squad as a combination of the features of the players it consists of. First, we just calculated average values of each skill for each team. This naive approach smoothes out each skill point; for example, there is only one goalie but other players" goal keeping skill make the difference small. Thus we need to sum up the stats according to the positions. To prevent smoothing out, we selected a few number of top scores from each squad. We tried several combinations of $2 \sim 5$ top scores for each feature. And we ended up selecting $[4, 5, 5, 5, 5, 4, 1]$ top scores respectively. Thus the one set of data looks as follows.

Team_Virtual_Features(team) =

$$\begin{cases} \sum \textit{top 4 Attacking,} \in [0, 2000] \\ \sum \textit{top 5 Skill, Movement,} \\ \quad \textit{Power, Mentality,} \in [0, 2500] \\ \sum \textit{top 4 Defending,} \in [0, 1200] \\ \sum \textit{top 1 Goalkeeping,} \in [0, 500] \end{cases}$$

An on-field player is represented as a set of 33 features which are split into 6 main categories Table I. A goalkeeper is represented as a set of 5 features. A squad is given as a set of 11 players, one of which is a goalkeeper the other 10 are usually split into a defensive group and a offensive group. The aggregation function for team is computed by taking all the goalkeeper features and averaging the field players features according by averaging top 4 Attacking and Defending, top 5 Skill, Movement, Power, Mentality features fromTable I The rationale behind this is that 1) goalkeepers features must be present in the aggregation because there is only one goalkeeper per team, 2) There are usually 4 defenders so it makes sense to only consider top 4 purely defensive skills. The other coefficients were approximated by trying a set of different combinations and choosing the one that produces best hit rates.

1) Match result: For each match, there are three outcomes: home team win, draw, and away team win. From the raw data, home team score and away team score, we converted them into binary values of three categories. For example, if home team score $>$ away team score, home team win=1 and rest of them are 0.

2) Feature selection: To find out which features are more relevant, we also conducted a feature selection over the data. We tried two feature selection algorithms: sequential forward selection, and best first selection. From the former, top 5 features were home team skill, home team movement, home team mentality, home team defence, home team goalkeeping. And from the latter, top 5 features were home team attacking , home team skill, home team movement, hometeam mentality, home team goalkeeping. With SVM, the accuracy of prediction from them was almost similar to ones that used the entire feature set.

IV. Models and Results(Supervised Learning)

We treated each match played between a home and an away team as a sample which is labeled as (home win, draw, away win). Notice that in soccer there are three possible outcomes for each game but most of our classifiers have binary outputs. We overcome this issue by slightly redefining the problem we are trying: we will have three binary classification problems where in each one we try to distinguish between on of the labels and the other two.

Example (for home team win prediction)

$$Y = \begin{cases} 0 & \textit{if home team won} \\ 1 & \textit{otherwise} \end{cases}$$

This classifier will basically tell us if the home team won or not.

A. Real Predictor

Each sample is a match between TeamA and TeamB which is labeled with Y. As feature set we used the Team_Real_Features(TeamA), Team_Real_Features(TeamB) which is a 24 dimensional vector. We applied Logistic Regression and Linear SVM to predict labeling of each match. We experimented with the feature set, by only looking at the teams performance in the last 3 matches in order to capture the "current form" of the team but the hit rate was not affected. We achieved hit rate of around .75(Table I) which is comparable to results in related works and is a good baseline for our "Virtual Predictor".

B. Virtual Predictor

Similar to the "Real Predictor" we combined the features for each line up, but we used the virtual features instead - {Team_Features_Virtual(TeamA), Team_Features_Virtual(TeamB)} which is a 66 dimensional vector. We applied Linear SVM, RBF SVM, Logistic Regression, SGD and Multivariate NB(we were required to discretize our values for this model). The "Virtual Predictor" produced results comparable to the "Real Predictor" and with a little bit of tuning Linear SVM performed better(Table 1).

V. Identifying Strategies(Unsupervised Learning)

Soccer managers are responsible for developing a team strategy before each match in order to surprise the opponent. Strategy includes things like player positioning, tactics, set pieces, etc. But team strategy is very often predicted by the sports-reporters by just examining the squad before the game. The reason this can be done is because managers trying to play an offensive strategy will have to include a lot of offensive players in the squad and visa versa. We take advantage of this property

TABLE II: Prediction results comparison

Real Data			
Model	Home Win	Draw	Away Win
Linear SVM	73%	75%	71%
Logistic Reg	73%	72%	74%
Virtual Data			
Model	Home Win	Draw	Away Win
Linear SVM	78%	80%	78%
RBF SVM	69%	81%	80%
Logistic Reg	70%	75%	76%
SGD	64%	70%	67%
Multinomial NB	78%	70%	75%

and attempt to use unsupervised machine learning techniques to identify different types of strategies based solely on the players skills present in the each squad. "Fixture results" collected from *http://www.goal.com* also includs the squads that were played in each match. We define an aggregation function similar to Team_Virtual_Rating, but notice that this function identifies strength of the team not the strategy. In order to identify strategy we need to normalize the features in Team_Virtual_Rating.

A. Preprocessing

Normalization is required for k-mean clustering, because Team_Virtual_Features for top teams are likely to be higher than the ones for weaker teams in every aspect - they will have better defense, offense, etc. But we normalize these features by making sure that all the scores in Team_Virtual_Features adds up to 1. Thus the value for each feature in Squad_Strategy will be represented in proportion to all the other features in the composition. This will allow us to observe if one of the features is overemphasized which will define the strategy. For example, a weak team and a very strong team may have different Team_Virtual_Ratings but they may have similar Squad_Strategy if they are playing a similar strategy.

Squad_Strategy= Normalized(Team_Virtual_Features)

Thus we can see that which skill is emphasized and which is not in a certain combination. We assume that this emphasis of skills reflects the strategy of the squad.

B. Clustering results

We computed Squad_Strategy for each squad played in the 380 matches of the season. We then applied k-means clustering to the data to identify distinct types of strategies with various number of k. Figure 1 shows the root mean square error of k-mean clustering for different number of k. Even though larger k shows better fit, if k is too large, it would be meaningless. From the RMSE plot, 5 or 6 will be appropriate number for k. Thus rest of analysis based on 5-mean clustering.

Figure2 shows the 5-mean clustering. Each cluster shows distinct combinations: cluster0 is well balanced except the attacking, cluster1 is more focus on attacking and individual skills, cluster2 is more biased on defending and movement, cluster3 is focusing on goalkeeping and individual skills, and cluster4 concentrates on goalkeeping and defending.

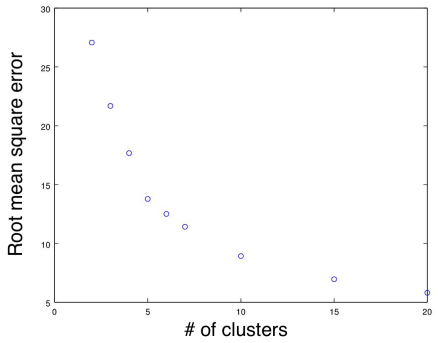


Fig. 2: K-mean RMSE

C. Clustering analysis

Given 5-mean clustering result, we analyzed winning odds between clusters. Hence we can see which combination, i.e. strategy, is more plausible in certain situations.

1) Winning odds between clusters: General winning odds between clusters are given in Table III. Winning odds are very different depending on the opponent and stadium. But in general, cluster4 does well. Since the table shows winning probability of home team, positive numbers are good for rows, and negative numbers are good for columns.

TABLE III: Winning probability of home team by clusters

Home\Away	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
Cluster0	0.00000	0.25000	0.40000	-0.75000	-0.75000
Cluster1	0.00000	0.33333	-0.50000	0.00000	0.20000
Cluster2	1.00000	0.00000	-0.33333	0.00000	-0.20000
Cluster3	0.50000	-0.20000	-0.66667	0.00000	0.00000
Cluster4	0.28571	0.50000	0.80000	0.14286	0.20000

However, Table III shows winning odds regardless of actual point differences. That means cluster4 may have more strong teams. Thus we also conducted analysis for weak teams; weak teams mean teams with less stat points than the opponent. TableIV shows winning cluster. This table shows cluster3 and cluster4 are good strategy for weak team in general: cluster3 can win against 1,2,and 3, and cluster4 can win against 0,1,and 3. Thus if a certain team is weaker than the opponent, it is better to focus on goalkeeping and defense to increase the winning odds like cluster3 and cluster4.

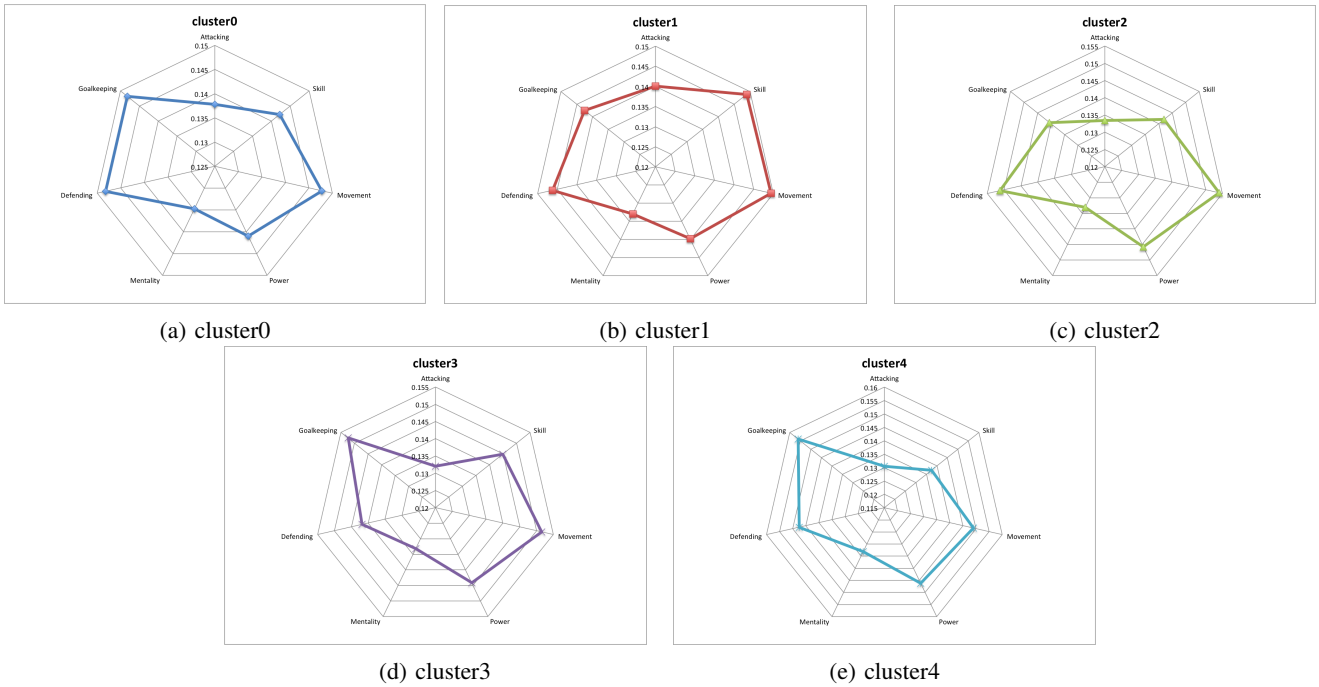
TABLE IV: Weak team's winning strategy

	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
Against	0	1	0,1	1,2,3	0,1,3

VI. Discussion

A. Supervised Learning

In this section we described two approaches to soccer match prediction: "Real Predictor" and "Virtual Predictor". The "Real Predictor" represents the traditional approach which applies machine learning to match statistics collected throughout the season(we refer to this as "Real Data"). This approach achieved



(a) cluster0 (b) cluster1 (c) cluster2

(d) cluster3 (e) cluster4

Fig. 3: Clusters. Each cluster represents a distinct strategy. Radial graph shows how different skills are emphasized in the each strategy.

0.75 hit rate for our data but we looked at related work and the highest hit rate achieved with this approach was 0.83[1]. "Virtual Predictor", which uses only data collected from video games, achieved 0.80 hit rate. This demonstrates that data collected by the video-game industry can be used to solve real problems in this medium, soccer, with comparable or even better results.

B. Unsupervised Learning

In this section we were able to use k-means clustering to identify 5 different types of playing strategies based on the players skills present in the squad(Figure 3). We were able to measure how these 5 different strategies perform against each other and observed that most top teams in Soccer have very offensive strategies. This is in contrast to NBA, where best teams where the ones with good defensive stats[1]. Also we observed that weaker teams perform better against stronger teams if they use defensive strategies.

VII. Conclusion and Future Work

One of the main challenges in machine learning projects is data collection which can be very time consuming and expensive. In this paper we demonstrate an alternative sources for curated data: video games. Video games are often overlooked due to its origin. However, video games have come a long way since Pac-Man and Frogger and have created phenomenally accurate simulations of the real world, which can only be done through very intensive data collection. This data can be used in machine learning projects to make predictions in the real world with very accurate results. Of course this would be made easier if the video game industry shared this information in public domain. FIFA 2015 has done a great job simulating the action of soccer players, but its likely that they have used more than 33 features, which are presented in the game, to accomplish this. We believe that if the entire feature set was available, "Virtual Predictor" would achieve even higher hit rates.

References

[1] M. Beckler, H. Wang, and M. Papamichael. Nba oracle. *Zuletzt besucht am*, 17:2008–2009, 2013.

[2] S. Luckner, J. Schröder, and C. Slamka. On the forecast accuracy of sports prediction markets. In *Negotiation, Auctions, and Market Engineering*, pages 227–234. Springer, 2008.

[3] A. McCabe and J. Trevathan. Artificial intelligence in sports prediction. In *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*, pages 1194–1197. IEEE, 2008.

[4] M. Spann and B. Skiera. Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1):55–72, 2009.