Football Prediction Using Poison Distribution

QUANG TIEN NGUYEN VAN
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU-HCM
Email: nvqtien@apcs.vn

THINH TU DUC
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU-HCM
Email: tdthinh@apcs.vn

THANG NGUYEN DUC
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU-HCM
Email: ndthang@apcs.vn

NGUYEN NGUYEN LE
Advanced Program in Computer Science
Faculty of Information Technology
University of Science, VNU-HCM
Email: nlnguyen@apcs.vn

***Abstract*—In professional sports, being able to know which team or individual will be the winner has always intrigue us for years, since the outcomes will result in the championships, teams qualify for tournament, etc. In this paper, we present methods which are simple and low computation cost, easy to implementation with a proper predictive result. Based on probability of events for a Poisson distribution, we generate the average goals each team can score each match, regard less of any other factors including their opponent. The result, tested with data of season 2017-2018 English Premier League and half way of the 2018-2019, with certain constraints, is stable at about 60 percent of match cases.**

***Index Terms*—professional sports, FIFA, football, EPL.**

I. Introduction

Football or soccer is regarded as the most popular sport in the world [2]. Two teams participate in a football match with the purpose is to score into the opponent goal. Whoever has more goals wins the match. In the case of 2 sides having the same goals when the match ends - usually after 90 minutes - a draw occurs. In several countries there are many football clubs competing for regional and national championships in leagues. All over the world, countries hold professional leagues for teams to participate for the national championship and even continent or global championships.

For years, people have been interested in knowing which team or who will crown the championships at the end of the league or tournament. Also, the end result of matches are no apart form peoples interest. Out of the many football leagues, we decided upon the English Premier League (EPL), which is the worlds most watched league with a TV audience of 4.7 billion people [4]. In the Premier League teams play against each others exactly two (2) times per season, one at the home stadium and the other at the oppositions ground . Therefore, for each team there will be 19 opponents (with two results in a season) [10].

In our project, we will discussing performance of various methods/models, and analyzing our results.

The goal of this paper is to propose simple methods with a good-enough predictive quality, low-cost computation and easy for implementation for predicting the outcomes of a match. From there, we can tell which team will be the champion or qualify for others tournament. We conducted our test to the 2017-2018 English Premier League - EPL.

To perform our predictions we use the Probability of events for a Poisson distribution [3] formula directly so that we can estimate the average goals scored by each team. Then generate simple matches scorelines predictions for 380 matches of the EPL 2017-2018.

II. Related Work

In paper [5] they attempt to develop a model following a Poisson distribution to predict the results and the best betting selections of the 2016-17 Premier League. They applying a varieties of method such as applying Probability Distribution of Goals Scored and Goals Conceded for teams in the Premier League and Multi-Variable Poisson Distribution with Home/ Away Factor all of these was tested with data from season 2011-2012 to 2015-2016. Their results was stable at about 66% for all three models.

In paper [9] M.J. Maher used Poisson distributions to model goal scoring in matches and also used other distributions such as bivariate and univariate Poisson distributions with attacking and defensive scores to predict the final result of a match.

In paper [8] Lee, Karlis and Ntzoufras' works implies that there is a low correlation between the two opponents number of goals which they can score. When one team scores, there is some chances that the other may do the same. The increased speed in game play may lead to more of these opportunities.

Barnett and Hilditch considered several other factors to determine their impact on match outcomes. Their works on whether artificial fields gave home teams advantages [6]. Dixon and Robinson's works studied in the variations in the rate a home team scores compared to that of away team. The result shown that the two rates depend on times passes in a match and which team is trailing or leading that match [7].

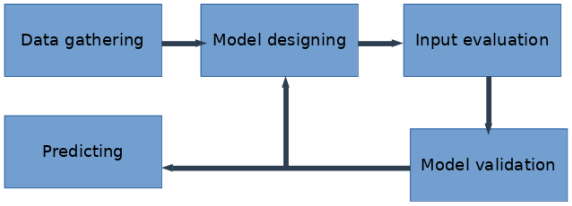

Fig. 1. Overview for our methods

III. METHODS

A. Overview

There are some few steps that we need to perform in order to construct each model. The very first step is to gather and select the data set which satisfies our needs. Then, we design the model for our prediction. After that, we need to evaluate the inputs in our model. After that, we run it through every match in our data set to validate how good is our model. If the result is not satisfies enough, we go back to our designing model step. And finally, we predict the results of the later half of the 2018-2019 season based on our constructed data sets.

There are many factors that effect the result of a match, but the objective of football is simple, score more goals then your opponent. Therefore, knowing how much goals are scored is critical to the way predicting the matchs outcome [3].

In order to accomplish this task, Following **Brianne Boldrins works** [4], we choose to use Poisson distribution. The Poisson distribution is popular for modeling the number of times an event occurs in an interval of time or space. An event can occur 0, 1, 2, times in an interval. The average number of events in an interval is designated λ (lambda). Lambda is the event rate.

In order to accomplish this task, we choose to use Poisson distribution. The Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space. An event can occur 0, 1, 2, times in an interval. The average number of events in an interval is designated λ (lambda). Lambda is the event rate.

$$P(\text{k events in intervals}) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

Where:

P(k) is the probability to have k goals in a match
λ is the average goals a team can score
e = 2.718 (Euler number)
k is our interest number of goals

Data was gathered from datahub.io [1]. We decide to use the EPL 2017-2018 data since from early observation its figures seems support our theory. From the data, we calculate the average goals a team can score.

A season usually take place from August until May of the next year and therefore a lot of things can change during this time. Managers be hired and get fired form the club. The signing of new players in the January transfer windows along with players leaving the club. All of these can heavily impact on the team performances and the average goals may or may not reflect exactly the team actual ability at any given time in the test period.

B. Game result prediction

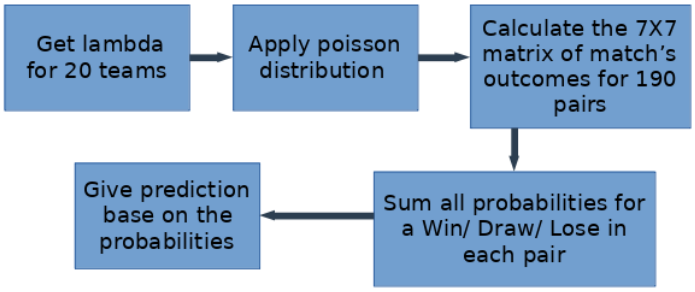

Fig. 2. Overview for Game result prediction

Fist we, need to get λ - which is team's average goals. This is simple as doing the division for 20 times. At this point, we finish the input evaluation step. Applying the poison distribution by plugging in the argument λ and calculate the probability of each team scores from 0 to 6 goals. Since the probability each team can score is considered to be independent, therefore using the product rule and we can form the 7x7 matrix of a pair which showing how likely is a specific scoreline to happen. And repeat this step for all the remains pairs. Notice that we can have our matrix cover more goals but in real world professional football, it is not likely that your team score about 5 or 6 goals in a match, even if you do manage to score that much goals, is it much like winning the lottery, the match is consider to have a shocking scoreline and it is extremely rare for it to happen.

Consider if we want to know team A's probability winning the match-up, summing all the scoreline's probabilities in which A scores more goals than B is fairly easy by using the matrix. The same concept can be applied for team A being defeat and getting a draw. In the end, we have 3 probabilities for team A to win, lose or draw (for team B is to lose, win or draw) and looking for which one of the three has the higher probability, that is our prediction for the match-up between the two teams A and B.

	Team B							
Team A		0	1	2	3	4	5	6
	0	P(0,0)	P(0,1)	P(0,2)	P(0,3)	P(0,4)	P(0, 5)	P(0,6)
	1	P(1,0)	P(1,1)	P(1,2)	P(1,3)	P(1,4)	P(1,5)	P(1,6)
	2	...	...	...	...	...	...	...
	3	...	...	...	...	...	...	...
	4	...	...	...	...	...	...	...
	5	...	...	...	...	...	...	...
	6	...	...	...	...	...	...	...

Fig. 3. How the outcomes matrix looks like

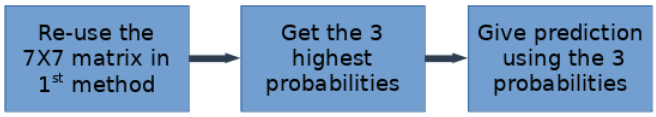

Fig. 4. Overview for 2nd method

C. Scoreline prediction

Our aforementioned method do comes up with a drawback. If we look back how our prediction computed, the draw outcome is only comprise 7 entries from the matrix while for the remaining two is 21 for each. Because of that, the most possible outcome would not likely to be a draw. So if the actual result is a draw, by default our prediction is wrong. In order to solve this, we decide only get the 3 highest possible outcomes in the outcomes matrix we did earlier on the 1st method. And now, we have 2 options, we can use the predictive result just to predict if the outcomes has a winning team or is a tie like what we did in the 1st one. Or we can use the predictive scoreline and check whether the actual result yields the same figures.

D. GiGD

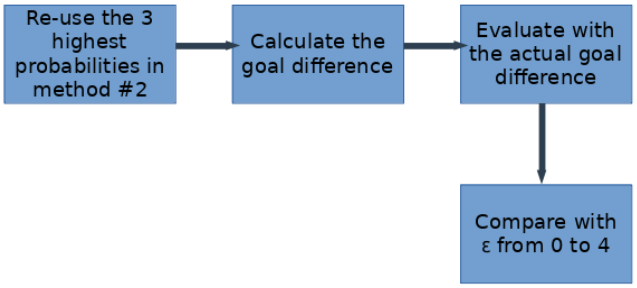

Fig. 5. Overview for GiGD

The concept of gap in goal difference (GiGD) is simple

$$\Delta G = GD_{actual} - GD_{predict} \tag{1}$$

Where:
ΔG *is gap in goal difference*
GD_{actual} *is Goal difference in actual result*
$GD_{predict}$ *is Goal difference in predictive result*

Now well talk about how this concept can help us to improve our accuracy. Look back at our scoreline method, we have a total of 36 match outcomes and our methods result has probability to be corrected comparing to the actual scoreline results is only about 1/36. By using gap in goal difference, we cover more entries in the outcomes matrix and therefore our result will be better.

But in terms of meaning of the results, this method can never be as good as the scoreline method. The Goal Difference itself can tell us who is our favor if we looking at its sign. If it is positive, that means the left hand side team is likely to beat the right hand side team and the opposite is true for the negative sign. Since this method is loss-tolerate when you accept your result even if it was wrong but only in some of our acceptable margin which we will discuss right after this.

Our goal still trying to get the score line as close as it can get. Reusing the result from our model 2, let ϵ be our **maximum accepted** gap in goal difference ranging from 0 to 4. And if we get our ϵ too big, the result is likely meaningless whatsoever. Consider if we predict an outcome to be 1-0 that means the goal difference = 1, and the actual result is 6-0 which means the goal difference = 6, that make our gap result in 5. so if we choose to let our ϵ accept this kind of result we will miss the goal of this model.

From **the 3 highest scoreline probabilities** we can calculate the Goal Difference in our 1st, 2nd and 3rd predictive result . We also need to get our Goal Difference for the actual result. Then our gap can be computed easily. We then proceed evaluating these gaps with ϵ. If our gaps is smaller or equal to the current interest ϵ its a pass, otherwise is a no.

IV. RESULT

The Game result prediction method give us 199 correct matches over a total of 380 matches in the season 2018-2019 resulting in a 52.37% of accuracy. But the problem considered before in the methodology section is that it barely give a tie prediction. To clarify, only two pairs between Huddersfield vs Swansea and Swansea vs West Brom gave out a tie prediction in which two of three teams were relegated at the end of the season. If we look further into our results, we have the higher number of success for the top six teams in our prediction than the remaining ones, which is illustrated in Figure **??**

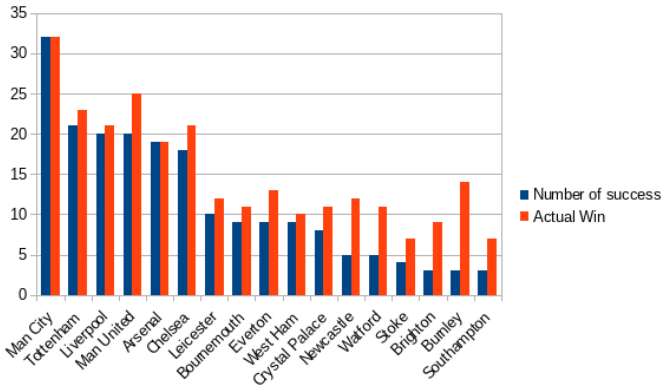

Fig. 6. Number of success vs teams actual win

Result of **scoreline prediction** method is 10.53%, 10.53%, and 11.31% respectively for 1st, 2nd, and 3rd highest percentage. These are not remarkable enough to meet our expectation, so we only focus on result of the improved version, which is the GiGD method. We have conclude the result of 3 highest percentage into 3 graphs, with x-axis be ϵ , and y-axis be the percentage of successful prediction

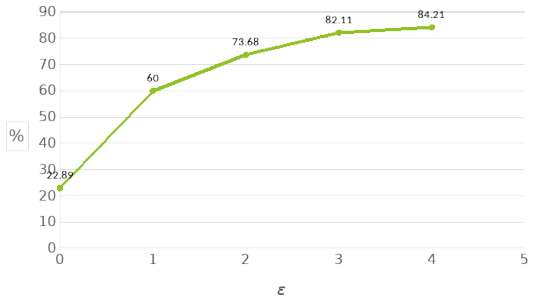

Fig. 7. Result for 1st predictive result

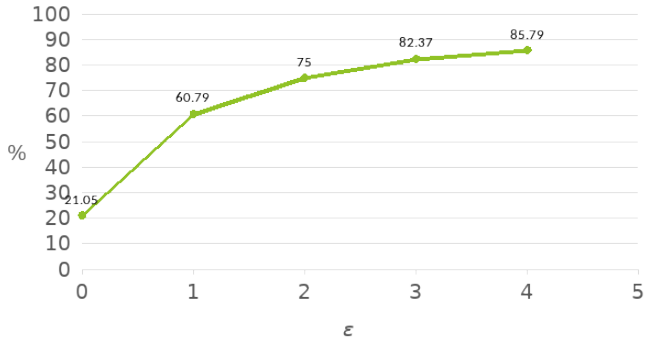

Fig. 8. Result for 2nd predictive result

As we can see from the 3 graphs, for $\epsilon = 0$, the results are 24.73%, 22.89%, and 21.05% respectively. Then, moving from $\epsilon = 0$ to $\epsilon = 1$, they increase nearly 40% more, to 62.63%, 60%, and 60.78%, which is the authors desired result. After that, results tend to increase less significantly for bigger and bigger epsilon. So the authors choose $\epsilon = 1$ because it hit the sweet spot for accuracy and the tolerability.

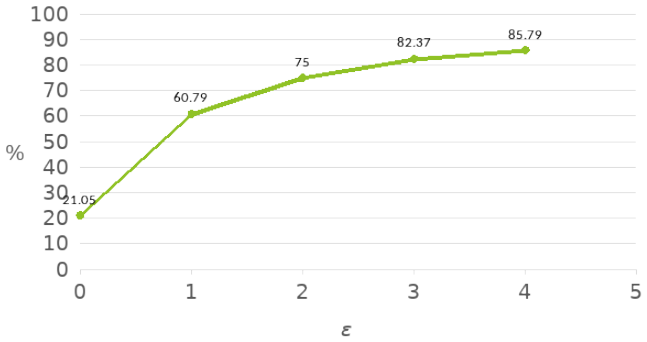

Fig. 9. result for 3rd predictive result

REFERENCES

[1] "English premier league datasets." [Online]. Available: https://datahub.io/sports-data/english-premier-league

[2] "The most popular sports in the world." [Online]. Available: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

[3] "Probability of events for a poisson distribution." [Online]. Available: https://en.wikipedia.org/wiki/Poisson_distribution

[4] "The worlds most watched league," 11 Dec 2014. [Online]. Available: https://www.premierleague.com/en-gb/about/the-worlds-most-watchedleague.html

[5] B. Boldrin, "Predicting the result of english premier league soccer games with the use of poisson models."

[6] Goddard and John, "Regression models for forecasting goals and match results in association football," *International Journal of Forecasting*, vol. 21, no. 2, pp. 331–340, 2005.

[7] H. Hamilton, "Goal scoring probability over the course of a football match — soccermetrics research, llc." [Online]. Available: soccermetrics.net

[8] D. Karlis, "Analysis of sports data by using bivariate poisson models," 2003.

[9] M. J. Maher, "Modelling association football scores," *Statistica Neerlandica*, vol. 36, no. 3, pp. 109–118, 1982.

[10] P. Marek and F. Vvra, "Home team advantage in english premier league."