Лаура Мухамадиева, Айман Молдагулова
Комбинированный метод распознавания лица

Аннотация. В настоящее время распознавание лица - одно из важнейших достижений биометрии, которое позволяет с помощью алгоритмов определять характер человека по его физиологическим характеристикам. Исследование и анализ существующих методов для распознавания лиц помогает распознать положительные и отрицательные аспекты каждого из предоставленных методов, на основе чего создается эффективный метод распознавания образов. Фундаментальной идеей данной статьи является улучшение консолидированной техники распознавания лиц с использованием библиотеки специализированного зрения OpenCV и вычисления Виолы-Джонса, с помощью которого лицо из видео-потока соотносится с определенными 68 точками лица. Дополнительно рассматривается расчет для естественного изменения различия изображения, чтобы продуктивно идентифицировать структуру лица. Кроме того, в работе удалось повысить качество распознавания объектов, снизить степень ложного подтверждения, сократить время подготовки классификатора и обработки изображения.

Ключевые слова: распознавание лиц, гибкие контуры, асимметрия лица, коррекция цвета.

Сведения об авторах:

Лаура Мухамадиева, PhD студент, лектор кафедры компьютерной инженерии и телекоммуникаций Международного университета информационных технологий

Молдагулова Айман Николаевна, к.ф.-м.н., и.о. профессора кафедры информационных систем.

About authors:

Laura Mukhamadiyeva, PhD student, lector, Computer Engineering and Telecommunication Department, International Information Technology University

Aiman N. Moldagulova, cand. of physics and math. sci., acting professor, Information Systems Department, International Information Technology University.

УДК 530.1, 681.3.06

Manabayev B.K.

International Information Technology University, Almaty, Kazakhstan

AN APPLICATION OF EXPLORATORY DATA ANALYSIS FOR EVALUATING FOOTBALL TEAM PERFORMANCE

Abstract. *The following article provides information on the usefulness of exploratory data analysis in identifying similar patterns and anomalies of how a football team performs throughout the whole season. This paper presents in-depth look on popular, controversial hypothesis while also describing a hidden shift in play-style of certain football squads, and what results have those teams achieved in accordance with data gathered by sports analytics web-resource.*

Key words: *exploratory data analysis, football team performance, expected goals metric, football statistics*

Introduction

Machine learning [1] is the science of computer algorithms which provide the capability to learn information from the given datasets without considering predetermined equations. Nowadays

one of the rapidly developing areas of machine learning is a predictive analytics, which can be further applied in scientific, industrial and business spheres along with sport domain.

However, it must be mentioned that each machine learning project, whether it is related to business domain or sports, usually requires one, yet important step before a data analyst starts to work with machine learning model. This step is called an exploratory data analysis (EDA) [2], which serves as an approach for visualizing, summarizing and comprehensively describing the crucial characteristics of the dataset. EDA provides means and tools for finding hidden insights and anomalies, while also transforming the given data in a way such that it could fit in future stages of building the machine learning model. Due to large set of available visuals such as charts, graphs, plots and heat-maps, EDA is ideally suited for investigating large sport datasets.

It is widely believed by many football enthusiasts that the strongest English squads from Premier League (also known as EPL) Big 5 have started to dramatically over-perform [3] in their home league, which plenty of fans connect to a smooth drop of other clubs' quality of play. Furthermore, 2 statements related to the proposed issue have been identified:

1) Squads from TOP5 EPL from year to year demonstrate a sharp increase of quantitative gap between strong football clubs and middle tier/outsideers. Thus, TOP5 EPL teams seem to over-perform frequently starting from season 2014/15;

2) Squads from the remaining TOP15 EPL are either existing in the middle of the table or compete with 2 random teams at worst in order to not be relegated to the EFL Championship. In other words, middle tier/outsideers experience a dramatic under-performance [3].

Consequently, the goal of this article is to find out whether the level of best football teams in England has increased significantly, or the average level of middle and low tier English football clubs is not enough to constantly compete with stronger squads.

In order to prove or disprove the statements mentioned above Microsoft Power BI Desktop application was chosen, because it lets a user to clean, build and shape visuals on the basis of multiple data sources. There are many types of visualizations presented in Power BI Desktop and most of them are ideally suited for exploratory data analysis, since visualization graphically represents the information from the dataset. Moreover, Power BI Desktop provides an opportunity to create a collection of visualizations in a single file called *a report*.

Data analysis and exploration

First of all, the dataset created by Kaggle user Sergi Lehkyi named "Football Data: Expected Goals and Other Metrics" was selected, which consists of detailed information about Top European Leagues advanced statistics starting from season 2014/15 till 2018/19. In order to implement the dataset Sergi has scrapped available summary information from the web-site <https://understat.com/> [4] to precisely look at performance of all football squads in numbers. Currently the dataset contains an overall statistics by the end of each season for 6 UEFA Football leagues such as La Liga, English Premier League, Bundesliga, Serie A, Ligue 1 and Russian Football Premier League.

The line graph below shows information about a difference in scored and expected goals sorted by football squads. Quadratic blocks on the right side of the graph demonstrate an amount of matches completed by the end of season. It is worth to mention that expected goals metric [5] serves for measuring a quality of how certain football team has been performing throughout a season.

Figure 1 – Comparison of scored and expected goals by football squads in season 2016/17

Looking attentively into the goal scored/expected goals graph of various time periods there is no doubt that in terms of amount of given parameters there are very few cases, which can be characterized as an obvious over-performance. For the 5 year period only Tottenham from London and Liverpool from Merseyside have presented a stable yet not always sharp over-performance in comparison with other clubs.

For example, in seasons 2014 and 2015 The Spurs goal difference [6] rate hasn't exceeded 5,61 goals. After that there was a dramatic increase in the difference rate, which had tripled and by the end of the season 2016 it has increased approximately up to 16 goals, which is actually one of the most significant goal differentials within 5 year period. It must be mentioned that Tottenham presented no over-performance signs in next seasons, which might be an indication of players' tiredness after outstanding performance in 2016. 2 years later Tottenham at best only scored 5 more goals than it was predicted by expected goals metric.

Some may also highlight Liverpool, which has been performing good enough to reach high positions in the Premier League. However, it is worth to note only seasons 2015 and 2016, at which The Reds scored 8,63 and 11,37 goals respectively. Such difference in amount of scored and expected goals is anomaly, yet the rate of Liverpool is not as significant as in case of London squad.

On top of that, it is important to mention that outstanding work on the field can be measured not only by the difference in expected goals and real scored ones, but considering PPDA [7] coefficient as well. To be precise, PPDA index shows the willingness of a particular squad to engage in

pressure on enemy's half. The less PPDA index is, the more aggressively does a certain team play. For instance, let's take a look on a PPDA rate for a season 2017/18.

Figure 2 – PPDA indices for different EPL teams in season 2018/19

The lowest PPDA rates belong to Man City, Arsenal and Tottenham squads. Citizens' numbers are quite impressive, because City players only allow ~6-7 passes on the opposition half, before they attempt a defensive move. By contrast, Tottenham players allow to do 1 more pass (their index is 8,05) and then Spurs trigger the pressure in order to counter-attack their opponent. Arsenal is only slightly worse than these 2 pressure leaders, having PPDA index of 9,58 that is still a great result in comparison with PPDA rates of other EPL squads.

An overall recession of pressure intensity between 2014/15 and 2018/19 seasons is shown on Figure 3 below.

Figure 3 – PPDA index has increased between 2014 and 2018, which only means that professional football clubs are shifting away from high intensity pressure on the field and might concentrate on improvements and innovations (like engaging in pressure by trigger [10]).

Along with PPDA coefficient, there is also a DC metric [8] to measure team performance. DC (also known as “deep passes completed”) is a quality metric that shows the proficiency of certain team in positional attacking. The higher DC rate, the better team X performs in terms of delivering the ball into opponent's penalty area. For instance, Liverpool has DC index of 431, which means LFC players, while having fewer passes in penalty area of opponents, have scored only 6 goals less than 18/19 EPL champion Manchester City, who made 582 passes at the same conditions.

It would be unfair not to mention how statistically awesome has been Tottenham between 2014 and 2018 years. For instance, in 2018 Tottenham was only allowed to make 270 passes within 18 yards for a goal, and Spurs scored 67 goals by the end of 18/19 season. Considering Arsenal FC, it is understood that Gunners made 130 passes more than Tottenham players did, but only scored 73 (!) goals. The difference is only 6 goals. Spurs undoubtedly had shockingly incredible rate of successfully executed scoring moments. These aspects of brilliant play by Tottenham originate generally from seasons 2014/15 and 2015/16, at which Spurs also haven't had much of passes allowed by opponents, but still gained their position in EPL Top-club list, thanks to great usage of scoring moments.

Figure 4 – Spurs scored 69 goals being limited only to 255 passes before opponent attempts defensive action. By contrast, Manchester City scored 71 but required 379 passes within penalty area (18-meters).

A stable over-performance of Tottenham actually has its roots in the ability of a squad to effectively use scoring chances with less effort in terms of amount of passes required.

Figure 5 – Tottenham card showing an increase in PPDA rate (it indicates that Spurs' pressure intensity has been reduced slightly) and a noticeable rise of missed goals

Another method of identifying under-performing or over-performing squads is visualizing the difference between actual points earned by certain team and expected points [9].

Table 1 – The only squads that are mentioned are the ones that have actual and expected points differential more than 6. These teams can be considered as playing too well

Season	Squad	True PTS	Expected PTS	PD
2014 (3)	Chelsea	87	75,32	11,68
	MU	70	63,03	6,97
	Swansea	56	43,32	12,68
2015 (3)	Leicester	81	68,94	12,06
	MU	66	56,44	9,56
	West Ham	62	49,8	12,2
2016 (4)	Chelsea	93	75,74	17,26
	Tottenham	86	75,37	10,63
	Liverpool	76	69,83	6,17
	Arsenal	75	62,12	12,88
2017 (3)	MC	100	91,09	8,91
	MU	81	62,33	18,67
	Burnley	54	41	13
2018 (5)	MC	98	90,64	7,36
	Liverpool	97	83,45	13,55
	Tottenham	71	61,44	9,56
	Arsenal	70	58,97	11,03
	West Ham	52	43,72	8,28

In fact, multiple clubs in EPL have a tendency to earn more points than other opponents. Noticeable signs of over-performance were spotted in case of Chelsea FC at seasons 2014 and 2016, but the squad's results have returned to a normal state (at least on the end of 2018/19).

Liverpool FC's difference between earned points and expected ones also became larger (from 6,17 in 2016 to 13,55 in 2018). Consequently, slight gap increase, which started in 2016, by the end of 2018 became much larger, indicating that Liverpool is either pretty luck squad or the team has been shifted towards team spirit and "last-minute" victories.

The most frequent cases in that comparison are West Ham – The Irons, and Manchester United. Despite having "one-time" clubs such as Swansea in 2014 and Burnley in 2017, a solid squad from London demonstrates the best results among middle tier/outsider teams in EPL. For instance, in 2015 The Irons have earned 12 points more than the model has predicted them to obtain. In 2018 the rate has reduced to 8.28 point gap.

On the other hand, Manchester United showed an undeniable upward trend. For example, in 2014 their PD rate was at 6,97 (almost 7) points. In the next year it increased on 2 and 2017 difference between expected/actual points achieved 18,67. This is the biggest gap for 5-year period among all clubs in EPL.

Overall, only 3 three clubs from EPL Big 5 presented very solid results in terms of over-performance by earned points. Those are Liverpool (hardly could be called a top-club considering later 2010s), Manchester United and Tottenham.

Expected points metric is more centered on showing the real difference between predicted values and final amount of points earned by EPL teams at the end of particular season. An example

below comprehensively describes which squads are doing too well, which ones have normal performing and which are seriously under-performing.

Figure 6 – Bar graph indicates MC, MU, Burnley and West Ham as the over-performing ones (earned points are painted in green), while teams such as Crystal Palace, (partially) Leicester, Everton and West Bromwich Albion obtained less points than expected points metric predicted them to

Conclusion

To sum up all the given information, it is clearly seen that the first statement is only partially true for very few top-clubs (like Tottenham and Manchester United), which tend to over-perform more frequently and obtain more points respectively.

However, not only TOP5 EPL teams, but also middle tier and even some of outsiders showed that tendency for performing way better than expected is not fully dependant on whether the club is top tier team or not.

The second statement “15 teams outside of Big 5 seem to under-perform for the last years” is disproved, because in case of comparison of parameters like goals, points, pressure and positional attacking, some squads from the bottom of English Premier League demonstrate a wide gap between predicted and real values. For example, West Ham among all middle-tier clubs in England showed the greatest gaps on comparative graphs. Other similar clubs and outsiders while presenting slight over-performing rates, still prove that even a low tier squad may showcase pretty solid results in the long run.

REFERENCES

1. Daumé III H. A course in machine learning //Publisher, cimpl. info. – 2012. – Т. 5. – С. 69.
2. Tukey J. W. Exploratory data analysis. – 1977. – Т. 2. – С. 131-160.
3. Ekstrand J., Waldén M., Hägglund M. A congested football calendar and the wellbeing of players: correlation between match exposure of European footballers before the World Cup 2002 and their injuries and performances during that World Cup //British journal of sports medicine. – 2004. – Т. 38. – №. 4. – С. 493-497.
4. <https://understat.com/league/EPL>
5. Pollard R., Ensum J., Taylor S. Estimating the probability of a shot resulting in a goal: The effects of distance angle and space //International Journal of Soccer and Science. – 2004. – Т. 2. – №. 1. – С. 50-55.
6. <https://web.archive.org/web/20140429133656/http://www.premierleague.com/en-gb/fans/faqs/who-wins-league-if-teams-finish-on-same-points.html>.
7. Robberechts P. Valuing the Art of Pressing //StatsBomb Innovation in Football Conference. – 2019.
8. <https://statsbomb.com/2015/07/converting-dangerous-passing-into-shots/>
9. Goldner K. A Markov model of football: using stochastic processes to model a football drive //Journal of Quantitative Analysis in Sports. – 2012. – Т. 8. – №. 1.
10. <https://themastermindsite.com/2019/04/11/how-to-coach-pressing/>

Манабаев Б.Қ.

Футбол командасының ойын сапасын бағалау үшін деректерді барлау талдауын қолдану

Андатпа. Бұл мақалада маусым бойы футбол командаларының ойынындағы ұқсас заңдылықтар мен ауытқуларды анықтау үшін деректерді зерттеудің пайдалылығы туралы ақпарат берілген. Осыған орай мақалада танымал, даулы гипотезаға терең көзқарас ұсынылған, сонымен қатар белгілі бір футбол командаларының қалай ойнайтындығы және спорттық-аналитикалық ресурста жиналған мәліметтерге сәйкес командалардың қандай нәтижелерге қол жеткізгені туралы жасырын өзгеріс сипатталған.

Түйінді сөздер: деректерді барлау, футбол командасының ойын сапасы, күтілетін голдардың метрикасы, футбол статистикасы

Манабаев Б.Қ.

Применение разведочного анализа данных для оценки качества игры футбольной команды

Аннотация. В данной статье рассматривается применение разведочного анализа данных для выявления сходных закономерностей и аномалий выступления футбольной команды на протяжении всего сезона. В статье представлен углубленный взгляд на популярную, противоречивую гипотезу, а также описан скрытый переход в стиле игры определенных футбольных команд, и то, каких результатов они достигли в соответствии с информацией, собранной спортивно-аналитическим ресурсом.

Ключевые слова: разведочный анализ данных, выступление футбольной команды, метрика ожидаемых голов, футбольная статистика

Авторлар туралы мәліметтер:

Манабаев Бекет Қанатұлы, «Компьютерлік инженерия және ақпараттық қауіпсіздік» кафедрасының магистранты, Бағдарламалық Инженерия ББ, ХАТУ

Сведения об авторах:

Манабаев Бекет Қанатұлы, магистрант кафедры «Компьютерная инженерия и информационная безопасность», ОП Программная инженерия, МУИТ