주성분 분석 및 K-Means 클러스터링을 통한 축구 선수 스카우팅 시스템

최동훈
아주대학교 소프트웨어학과, dhc4003@ajou.ac.kr

요약

데이터 과학과 결합하여 괄목할 만한 성장을 이루고 있는 스포츠 산업이지만 선수를 영입하는 기준이 애매모호한 경우가 많다. FIFA23 데이터에는 전 세계 축구선수들의 다양한 능력치가 있는데 이를 주성분 분석을 통해 차원 축소를 진행하고 K-means 클러스터링을 통해 알맞은 군집 분석을 수행한다. 이를 다양한 시각화 방법을 활용하여 각 군집의 특성을 파악한 뒤에 같은 군집 내에 코사인 유사도를 구하여 이 수치가 높은 선수를 알려준다. 즉, 팀 내 핵심 선수의 대체자를 이와 같은 데이터 마이닝 알고리즘을 활용하여 객관적이고 논리적으로 구할 수 있게 된다.

1. 연구 배경

전 세계적으로 스포츠 산업은 빅데이터(Big Data)와 AI 산업의 발전으로 인해 분석적이고 체계적인 방향으로 나아가고 있다. 2011년에 개봉한 영화 '머니볼(Moneyball)'은 단장 빌리 빈(William Lamar Beane)이 메이저리그 최하위 팀인 오클랜드 애슬레틱스(Oakland Athletics)를 통계학과 경제학을 접목한 데이터 분석을 통해 기적을 이뤄내는 영화이다.[1] 이처럼 빅데이터 분석은 이미 스포츠 산업에서 다양한 방식으로 적용되고 있으며 이를 스포츠 데이터 분석이라 부른다.

그중에서도 축구는 데이터 분석이 필수적인 종목이다. 과거부터 유명 구단들은 실력 있는 선수를 영입하기 위해서 거액의 돈을 아끼지 않는다.[2] 영국 BBC에 따르면 EPL(English Premier League) 클럽들이 올해 여름 이적시장에서 지출한 금액은 19억 파운드(약 3조 원) 정도다.[3] 하지만 잘못된 스카우트(Scouting)로 영입된 선수는 클럽에 막대한 손해를 끼친다. 선수의 경기 형태(playstyle)보다 직감이나 단순 수치(stats)로 선수를 영입하는 경우가 대부분이다. 이러한 영입 방식은 논리성과 객관성을 만족하기에는 어려움이 있다. 수많은 축구선수가 있고 이들의 능력치를 나타내는 지표들은 많아서 효과적으로 검색하고 처리하여 영입의 선택 기준이 될 필요가 있다.

이 목표를 달성하기 위해 주성분 분석(PCA:Principal Component Analysis), K-평균 클러스터링(K-means clustering), 유사도 분석 등의 데이터 마이닝(Data Mining) 알고리즘과 접목하여 선수 영입을 위한 모델/APP를 만든다. 이를 클럽에 적합한 선수를 찾기 위한 도구로 사용할 수 있다. 주성분 분석에서는 다변량 데이터를 분석 대상으로 삼아 분산을 가장 잘 설명하는 축의 개수를 선정해서 그 축에 따라 축소, 요약하면 그 데이터가 주성분이 된다. 이를 통해 데이터에서 나타나는 경향성으로부터 표면에 드러나지 않는 숨겨진 정보를 도출해 낼 수 있다. 즉, 공분산 행렬의 고윳값(eigenvalue)과 고유벡터(eigenvector)를 구하여 데이터 벡터를 어떤 벡터에 내적할 것인지에 대한 최적의 답을 구할 수 있다. 그 뒤에 K-평균 클러스터링을 사용하여 데이터를 K개의 군집으로 나눈 뒤 초기에 중심값(centroid)을 설정하고 데이터마다 가까운 중심값이 속해 있는 군집으로 할당한 뒤에 중심값의 위치가 변하지 않을 때까지 중심값을 군집의 중심으로 이동하고 데이터를 군집에 할당하는 과정을 반복한다.

2. 연구 내용

2.1. 아이디어 및 데이터 전처리

이 프로젝트를 위해 kaggle에 있는 FIFA23 dataset를 활용한다.[4] 이 데이터는 18,000명 이상의 선수와 90여 개 정도의 신체 능력치, 포지션별 능력치, 세부 능력치를 feature(특징)로 갖고 있다. 무엇보다도 선수의 재능을 발견하고 축구 클럽이 선수의 가치를 결정할 때 작용하는 특징을 스카우트 분석에 활용할 수 있다. FIFA23 data는 정제된 데이터이기 때문에 결측치가 없다. 다만 column name과 column 안의 데이터에 공백이 있을 수가 있어 없애는 작업(trim)을 한다. 이후 선수의 데이터를 골키퍼(GK: goalkeeper), 필드 플레이어(Non_GK: field player)로 분류한 후 선수마다 능력치를 나타내는 feature data만 30여 개나 되기 때문에 특징을 잘 나타낼 수 있는 7개의 데이터(pass, shoot, pace, skill, movement, defense, physical) 로 재분류한다.

<표 1> 표준화된 필드 플레이어 데이터 (7명)

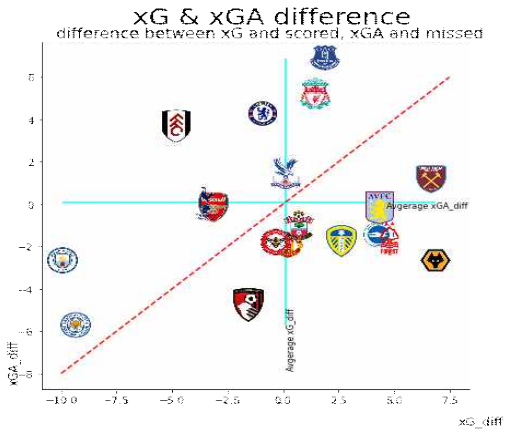
index	pass_index	shoot_index	pace_index	skill_index	movement_index	defense_index	physical_index
0	3.43	2.8	1.2	3.15	3.67	-1.22	0.77
1	2.45	2.67	1.02	2.65	3.09	-1.15	2.02
2	2.24	2.85	0.65	2.45	3.09	-0.85	2.82
3	3.52	2.8	0.56	2.65	2.86	0.67	1.2
4	2.14	2.58	2.62	2.95	3.49	-1.07	1.76
5	2.31	2.7	1.98	2.65	3.52	-0.4	1.45
6	2.22	2.93	1.15	2.4	3.35	-1.36	2.02

또한 실제 축구에서 사용되는 xG(기대 득점:Expected Goals), xGA(기대 실점:Expected Goals Against) 등의 지표를 가져오기 위해 Understat 사이트에서 유럽 최상위 리그에 대한 자세한 통계를 확인할 수 있다.[5] xG란 팀이나 플레이어가 시도한 슛의 질과 양을 바탕으로 예상되는 골 수이다. xG값이 0.5이면 득점 확률이 50%이며 xG값은 슛의 질을 결정하는 좋은 방법이다. 반대로 xGA란 시도한 슛의 질과 양을 바탕으로 실점할 것으로 예상되는 팀의 골 수이다. 위와 같은 예측 지표를 통해 단순히 결과만 보던 기존 분석 방식에서 벗어나 향후 구단의 성적을 예측해볼 수 있다.[6]

Understat의 데이터를 Web scraping을 통해 가져온다. EPL 팀들의 여러 통계 지표를 가져와 EPL_df라는 이름의 data frame을 만든다. 이 중 xG_diff는 xG와 실제 득점 수의 차로 이 값이 양의 방향으로 클수록 기대 득점에 비해 골을 넣지 못했으며 팀의 골 결정력이 낮다고 볼 수 있다. xGA_diff는 xGA와 실제 실점 수의 차로 이 값이 음의 방향으로 클수록 기대 실점에 비해 더 많은 실점을 기록했으며 팀의 수비력이 좋지 않다고 볼 수 있다.

이를 시각화하여 표현할 경우 그림 1에서 볼 수 있듯이 xGA_diff의 평균값을 x축, xG_diff의 평균값을 y축이라 가정했을 때, 제 사분면에 해당하는 팀들이 하위권일 확률이 높으며 실제로 중위권인 Brighton(Brighton & Hove Albion FC)을 제외하곤 강등권에 가까운 팀들임을 확인할 수 있다.

(그림 1) xG & xGA difference

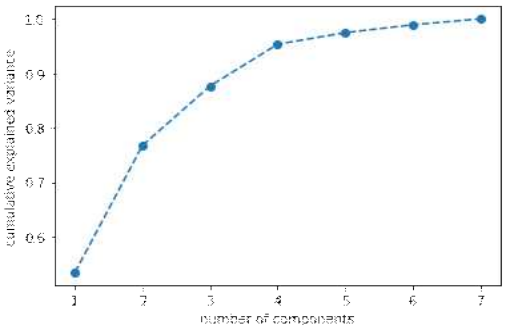


2.2. 주성분 분석(PCA)

본 논문에서는 FIFA23 데이터에 대해 주성분 분석을 수행하였다. 우선 앞서 필드 플레이어 데이터는 표준화(standardazation)를 통해 데이터 스케일링(data scaling)을 수행하여 분산량이 왜곡되는 것을 막는다. 그 후 7개의 선수 특성 관련 데이터를 차원 축소를 통해 줄이고자 한다. 아래의 그림2는 주성분 각각의 고윳값을 고윳값 전체를 더한 값으로 나눠 준 것이며 이를 통해 해당 주성분의 고윳값이 차지하는 비율을 알 수 있다. 또한 알고리즘을 통해 누적 분산량이 전체 중 95%를 넘을 때 기준으로 주성분 개수(n_components)를 4개로 결정하였

다. 즉, 4개의 주성분으로 전체 분산의 95% 이상을 설명할 수 있다는 뜻이다.[7]

(그림 2) PCA 누적 분산량

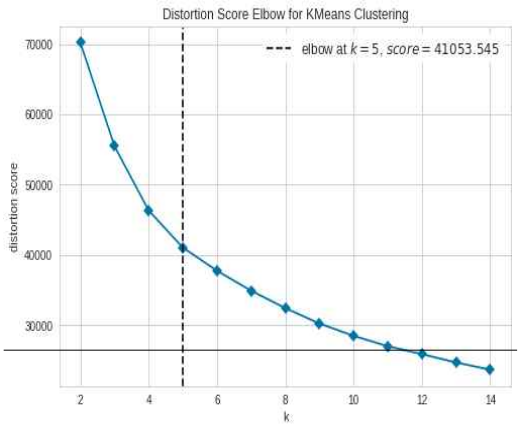


2.3. K-Means clustering

PCA를 통해 차원 축소한 데이터는 원래의 데이터에서 변형이 되었기 때문에 의미를 찾아내고 시각화하기 위해서는 K-Means 클러스터링이 필요하다.[8]

그림 3을 통해 보면 K-means는 미리 클러스터 수 k를 지정해야 하는데, 같은 군집 내 WCSS(Within Clusters Sum of Squares)가 급격히 완만해지는 구간의 k를 선택하는 'Elbow Method'를 사용했다.[9] 그림 3에서 제일 많이 구부러지는 구간이 5라고 판단, k 값을 5로 설정했다.

(그림 3) K를 찾기 위한 Elbow Method



이후 클러스터링을 진행한 뒤에 표2와 같이 PCA를 통해 얻은 값과 해당 cluster의 번호가 추가된 data frame을 얻을 수 있다. 테스트할 때마다 군집의 번호는 다르게 나오지만 군집에 소속된 선수는 같다.

〈표 2〉 주성분 분석, 클러스터링을 마친 후의 Dataframe

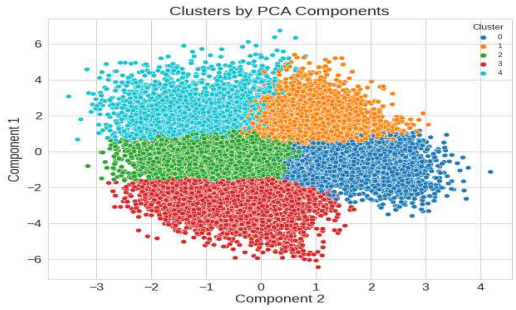
index	FullName	Cluster	BestPosition	PCA Component 1	PCA Component 2	PCA Component 3	PCA Component 4
0	Lionel Messi	4	CAM	6.726875	0.342237	0.615166	0.181881
1	Karim Benzema	4	CF	5.816280	-0.336014	0.132669	1.394052
2	Robert Lewandowski	4	ST	5.701664	-1.078143	0.189942	1.999844
3	Kevin De Bruyne	4	CM	5.878852	-1.407581	0.854530	-0.417365
4	Kylian Mbappé	4	ST	6.356858	0.249 2178	-1.293110	0.964180

이번 테스트를 기반으로 한 분류 결과를 대강 살펴보면 0번 군집은 포지션이 스트라이커인 공격수, 1번 군집은 포지션이 윙어와 같은 속도가 빠른 공격수, 2번, 3번 군집은 모두 수비수, 4번 군집은 우리가 흔히 아는 유명 선수들로 형성되어 있다.

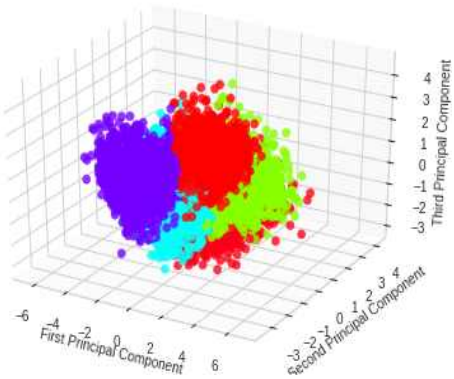
2.4. 시각화(Visualization)를 통한 평가(Evaluation)

군집 결과를 좀 더 뚜렷하게 보기 위해 시각화(Visualization)를 활용한다. 우선 2차원으로 살펴보면 그림4와 같다. 수비수가 포함된 군집이 PCA component 1,2 에서 모두 낮은 수치를 보인다는 것을 제외하면 선뜻 결과를 알기 어렵다. 이는 그림5와 같이 주성분을 3개로 늘려 3차원으로 보아도 분류 결과를 명확하게 해석하기 어렵다.

(그림 4) 2차원 시각화

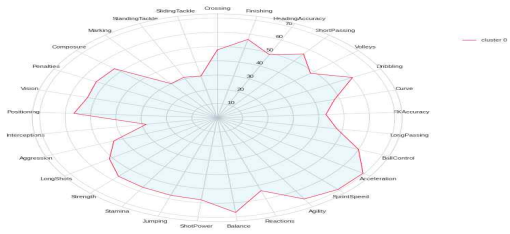


(그림 5) 3차원 시각화

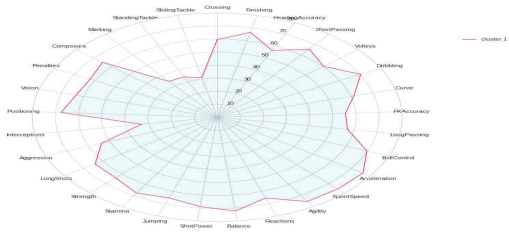


하지만 방사형 차트(radar chart)를 통해 시각화하고 데이터프레임의 선수들과 비교해보면 군집 분석 결과를 쉽게 설명할 수 있다. 군집마다 주성분 분석을 하기 전의 원래 필드 플레이어 Feature에 투영하여 특징을 정확하게 파악할 수 있다. 0번, 1번 군집은 거의 모양이 유사한 공격수 군집이다. 하지만 속도, 슛, 힘 등 전반적인 능력치가 1번 군집이 더 좋다. 2번, 3번 군집도 마찬가지로 서로 모양이 비슷한 수비수 군집이지만 2번 군집의 전반적인 능력치가 더 좋다. 마지막으로 4번 군집은 거의 모든 능력치가 골고루 좋으며 all-round player, 혹은 수준급 선수들의 군집으로 분류됨을 확인할 수 있다. 즉, 각 군집의 특성을 정확한 세부 능력치를 통해 파악할 수 있다.

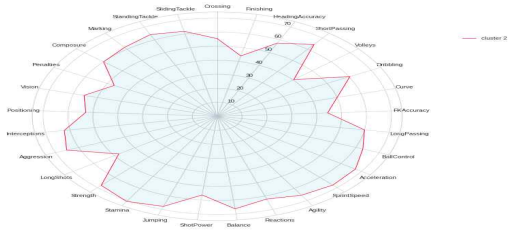
(그림 6) 0번 군집-공격수



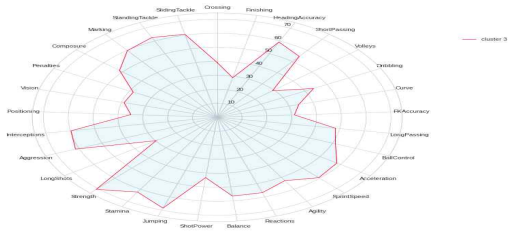
(그림 7) 1번 군집 - 공격수



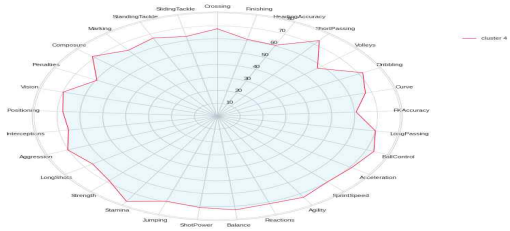
(그림 8) 2번 군집 - 수비수



(그림 9) 3번 군집 - 수비수



(그림 10) 4번 군집 - 육각형 선수



3. 기대효과 및 활용

앞서 군집화한 데이터들을 바탕으로 입력한 선수를 대체할 수 있을 만큼의 유사한 선수를 알려주는 프로그램을 만들 수 있다. 본 논문에서는 다양한 유사도 공식 중에 제일 보편적인 코사인 유사도 공식(Cosine similarity)에 따른 유사도 분석을 수행할 것이다. 코사인 유사도 공식은 그림11과 같다. 이 기법은 서로 다른 두 개의 벡터의 유사도를 측정하는 데 특화된 수식으로 클러스터링 연구 분야에서 많이 활용되는 기법이다. [10]

(그림 11) 코사인 유사도 공식

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

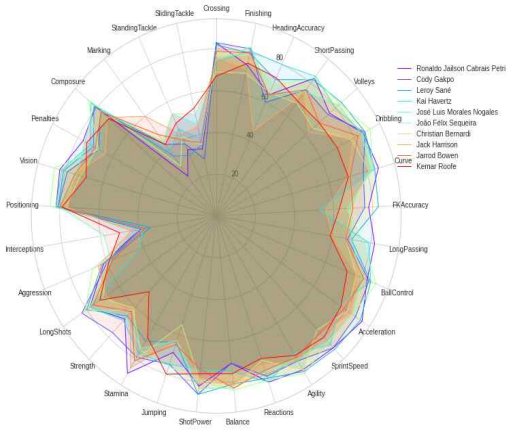
선수 이름(FullName)을 입력한 뒤에 코사인 유사도를 구하는 함수를 만들어 실행할 수 있다. 예를 들어 손흥민(Heung Min Son) 선수를 입력값으로 할 때 같은 군집 안에 있는 모든 데이터와 유사도를 계산하여 그 수치가 높은 순서대로 출력한다.

<표 3> 손흥민 선수와의 유사도 상위 10명

	FullName	Similarity
0	Ronaldo Jailson Cabrais Petri	0.99917
1	Cody Gakpo	0.99902
2	Leroy Sané	0.99871
3	Kai Havertz	0.99759
4	José Luis Morales Nogales	0.99744
5	João Félix Sequeira	0.99724
6	Christian Bernardi	0.99709
7	Jack Harrison	0.99682
8	Jarrod Bowen	0.99632
9	Kemar Roofe	0.99630

또한 이번엔 선수의 특성을 방사형 차트에 나타내는 알고리즘을 만들어 확인해 보면 그림11과 같이 선수와 유사한 능력치 혹은 그래프 모양을 가진 선수들이 similar player로 제대로 출력되는 것을 확인할 수 있다. 이들의 포지션(Position)은 모두 공격수이며 손흥민 선수처럼 빠른 속도를 가진 선수들이다. 이들의 손흥민 선수에 대한 유사도가 높게 나온 것을 보면 앞서 수행한 데이터 마이닝 알고리즘들이 정상적으로 실행되었음을 확인할 수 있다.

(그림 12) radar chart for similar player



4. 결론

xG, xGA, xpts와 같은 실제 축구에서 사용 중인 데이터를 통해 구단의 문제점을 파악하고 이에 대한 해결책을 게임 데이터를 통해 찾아 다양한 데이터들을 활용했다는 데 의의가 있다. 실제로 선수를 나타내는 데이터가 상당히 많아서 K-means 클러스터링을 바로 하지 않고 주성분 분석으로 적절히 데이터를 처리한 뒤에 클러스터링하여 더 좋은 군집화 성능을 보였다. 이를 바탕으로 유사도 분석까지 하여 선수들의 특성을 파악하고 대체자를 추천해 주는 알고리즘을 만들어 앞서 말한 구단의 문제점을 해결하는 방안을 만들었다고 생각한다. 다만 보완해야 할 점이 몇 가지가 있다. 본 논문은 필드 플레이어에 대해서만 데이터 마이닝을 수행하여서 골키퍼를 고려하지 않았다. 또한 알고리즘적 요소나 앱으로 구현한 요소가 부족했다고 생각한다. 마지막으로 본 논문은 구단에서 활약이 좋은 선수의 대체자 영입에 집중했지만 활약이 저조한 선수일 경우 이들의 능력치를 기반으로 유사도 분석하면 유의미한 결과를 얻기 어렵다. 따라서 해당 구단의 전술을 분석하여 그에 맞는 새로운 선수를 찾는 알고리즘을 추후 고안해볼 수 있다.

참고 문헌 (참고자료)

[1] 김윤후, 김상헌, 최형준, 정재은.(2018).빅데이터 분석과 게임이론을 활용한 야구선수 영입 모델. 한국컴퓨터정보학회 학술발표논문집 ,26(2),321-322.

[2] MICHAEL PARK, 이경묵.(2021).Liability of High Status: Overpayment to Relieve Status Anxiety in the English Premier League.Seoul Journal of Business,27(1),23-48.

[3] Transfer deadline day: Premier League spending reaches record £1.9bn for summer window . (2022). https://www.bbc.com/sport/football/62758471

[4] FIFA23 OFFICIAL DATASET . (2022).

Retrieved from

https://www.kaggle.com/datasets/bryanb/fifa-player-stats-database?select=FIFA23_official_data.csv

[5] EPL xG Table and Scorers for the 2022/2023 season .

(2022). https://understat.com/league/EPL/2022

[6] What is an expected goal? Description of expected goals . (2018). Retrieved from https://www.pinnacle.com/ko/betting-articles/Soccer/expected-goals-explained/B8Q2HGJ7XMJRZ58C.

[7] 서창우, 임영환.(2009).화자식별을 위한 전역 공분산에 기반한 주성분분석.말소리와 음성과학,1(1),69-73.

[8] Struyf, A., M. Hubert, and P. Rousseeuw, "Clustering in an Object-Oriented Environment," Journal of Statistical Software, Vol.1, No.4, pp.1-30, 1997.

[9] 박수연, 이도길.(2022).K-평균 클러스터링 및 주성분 분석을 활용한 뉴스 앱 이용자 유형 분류.한국정보과학회 학술발표논문집,(),1372-1374.

[10] https://en.wikipedia.org/wiki/Cosine_similarity