Bilge International Journal of Science and Technology Research
ISSN: 2651-401X
e-ISSN: 2651-4028
Research Article
2022, Volume: 7, Issue: 1, 43-48
Received: 14.04.2022; Accepted: 29.01.2023
DOI: 10.30516/bilgesci.1097014

Examination of Player Positions by Cluster Analysis

Okan Dağ[1*], Asım Sinan Yüksel[2], Şerafettin Atmaca[3]

Abstract: Today, the football industry stands out among the sports branches. Especially with the development of technology and its integration into football, different tactical understandings and formations emerge. With these developments, the current positions of the players and the other positions they are prone to play can be revealed as a result of the analysis. In this way, club management and technical team aim to establish the best team according to the current budget and tactical game understanding. Therefore, it is very important for the teams to play the players in the right position or to transfer the right player to the required position. In football competitions where 11 players are involved in the game, tactical changes can be made within the game according to the tactical arrangement and tactical understanding of the opposing team, and the player can be played in different positions. In this study, the player data of Turkey and the leagues of Germany, England, France, Spain, Italy, which are considered to be the five big leagues, for the years 2020-2021 were obtained from the website named "whoscored". In the data set obtained, the players who stayed on the field for a minimum of 1500 minutes were taken as a basis and clustering analysis was performed with the data of 985 players. Players are clustered on four basic positions: goalkeeper, defender, midfielder and attacker. In the study, Expectation Maximization, one of the clustering analysis algorithms, was used and a success rate of 81 percent was achieved.

Keywords: Artificial Intelligence, Data Mining, Clustering.

[1]**Address:** Suleyman Demirel University, Faculty of Economics and Administrative Sciences, Isparta/Türkiye.
[2]**Address:** Süleyman Demirel University, Engineering Faculty, Isparta/Türkiye.
[3]**Address:** Isparta University of Applied Sciences, Isparta/Türkiye.

***Corresponding author**: okandaq@gmail.com

Citation: Dağ, O., Yüksel, S. A., Atmaca, Ş. (2023). Examination of Player Position by Cluster Analysis. Bilge International Journal of Science and Technology Research, 7(1); 43-48.

1. INTRODUCTION

Sports have an important place in the social life of today's societies. Sports activities, which develop day by day, have become a phenomenon that affects education, economy and social structure in the 21st century (Yetim, 2000).

One of the sports branches that increase the impact on human life day by day is football (Öntürk, et al., 2019). Football, which has increased its visibility with the development of technology, brings entertainment, sadness, joy, national and international achievements as well as management, technical team and licensed players (Talimciler, 2008).

Football, which started with the first match played between Sheffield and London in 1966, managed to drag the masses after it and continued its development until today. According to official records, the first football match on Turkish soil was played in Thessaloniki in 1875 (İnan, 2007). In the globalizing world, football has succeeded in creating its own economy by separating it from other sports branches with its fan base, advertising revenues, sales of team products, transfers and viewing (Aygün & Ulucenk, 2019). The main sources of income that make up this economy are; box office revenues, athlete transfers and rentals, membership revenues, donations and aids, broadcasting rights revenues, advertising revenues, sponsor revenues (İnan, 2007).

Behind the football competitions in which two teams of 11 people take place on the field, there are elements such as management, technical team and fans. The team structure planned with the right transfers is important for football teams who want to have a good season. The current management and technical team aims to establish the best team and achieve the best results, taking into account the current budget. The management and technical team, by determining the transfer policies in this direction, want to transfer the players in accordance with the game plan of the

team, to strengthen the team tactically and to bring the players to the team that will excite the fans. With the staff established at the end of the transfer period, the technical team aims to increase the quality of the team by creating a player group that can play in various tactical formations such as 4-4-2, 4-3-3, 3-5-2, and trainings such as attack, defense, technique and individual. It is very important for the teams to play the players in the right position within the squad planning or to transfer the right player to the needed position.

Football data is of great importance today. Many analyzes such as data mining are used to reveal meaningful information from these large data sets. With the developing technological innovations, it has become easier to store data and large-scale data has become available. The analysis of data with human skills has created an inverse proportion with the increasing amount of data, and the analysis of big data has become difficult. Computer technologies have been used to make the raw data information or meaningful as a result of the analysis, and data mining analysis methods have been developed. Data mining includes disciplines such as data visualization, artificial learning, and statistics. Data Mining models, which include predictive and descriptive models, are grouped into three main groups: Classification, Clustering and Association Rules, according to the functions they perform (Savaş, et al., 2012). Cluster analysis is a multivariate statistical analysis method that enables grouping individuals, units or objects according to their basic attributes. In this method, individuals, units or objects that are similar to each other are grouped in the same group according to the selected selection criteria. As a result of the analysis, while the data in the same cluster are similar, they are not similar to the data in the other cluster (Kangalli, 2014).

In their study in 2019, Narizuka and Yamazaki developed a clustering algorithm that enables to create a multi-team formation using the Delaunay method, which defines a team's in-game formation as the adjacent matrix of Delaunay triangulation (Narizuka & Yamazaki, 2019). Kawasaki et al., in their study in 2019, proposed a new method to create a passing net based on the measurement of passing passes in the game. In 2016 and 2017, 9 matches of Fagiano Okayama team in Japan Professional Football 2nd League were analyzed. The rust network was divided into clusters by the clustering method and the rust network between different clusters was formed. It has been abstractly demonstrated that the developed passing network can pass successfully (Kawasaki, et al., 2019). Iman Bahrevan and Sayed Mohammed Ravazi predicted the player transfer market value with 74% accuracy by using machine learning methods in their study in 2021 (Behravan &Razavi, 2021). Johannes Stübinger and her friends conducted a study in 2019 on the prediction of match results using machine learning methods based on player characteristics. The study, which included the data set of Europe's five big leagues between 2006 and 2018, obtained statistically and economically significant results (Stübinger, et al., 2020).

In this study, the data of 985 players who played in the leagues of Turkey and Germany, England, France, Spain and Italy, which are considered to be the five big leagues, and took a minimum of 1500 minutes in the 2020-2021 season, were taken from the website called "whoscored". In line with these data, it is aimed to cluster the football players considered according to their goalkeeper, defender, midfield and attacking positions by using the Expectation Maximization algorithm.

In the study, real data were used; The clustering of the football players according to four basic positions was carried out. Thanks to the results obtained as a result of the analysis, alternative positions where the players can be useful besides their original positions have been revealed.

2. MATERIAL AND METHOD

There are many different definitions for data mining. Data mining is the process of extracting valuable information from large-scale data. In this way, it is possible to reveal the confidential relationships between the data and to make forward-looking predictions when necessary. In other words, data mining can be defined as the process of obtaining meaningful information from large data sets.

The data mining process is similar to classical statistical applications. However, classical statistical techniques are mostly applied on regular and special data sets. In the data mining process, studies are carried out on very large data groups and features (Özkan, 2016).

Techniques used in the data mining process;

- Classification,
- Clustering,
- Association Rules

grouped under three main headings. Among these techniques, classification is among the predictive models, while clustering and association rules are among the descriptive models (Gemici, 2012).

In this study, cluster analysis, which is one of the descriptive models, and Expectation Maximization (EM), which is one of the cluster analysis algorithms, will be emphasized.

Cluster Analysis

Cluster analysis is a type of analysis that ensures that similar elements in the data set are included in the same clusters. In the cluster analysis, the elements in the same cluster are as homogeneous as possible; It is desired that the elements in different clusters be as heterogeneous as possible (Akpınar, 2014).

Looking at the literature, there are many different types of clustering techniques related to cluster analysis. It is possible to collect clustering techniques under five main headings. These; They are grid-based, hierarchical, density-based, probability-based, and partitional clustering methods (Choudhary, 2016).

In the study, many algorithms were tried in the clustering analysis made by using the football player data obtained from the whoscored website, and it was seen that the Expectation Maximization (EM) algorithm gave the most appropriate result. The EM algorithm used in the study is given in detail in the next step.

2.1.1. Expectation Maximization (EM) Algorithm

The EM algorithm is one of the clustering algorithms that aims to find the maximum likelihood solution for hidden variable models. The EM algorithm can be defined as an optimization algorithm based on iteration (Dempster, 1977).

The algorithm performs the basic steps iteratively until it reaches the best convergence in terms of working principle. These steps, which are applied iteratively, are two, namely the E-step and the M-step. The E-step is called the expectation stage, while the M-step is defined as the maximization step. In the e-step, first a random probability distribution function is created, then the logarithmic probability function is found by obtaining the logarithm of the current function. Upon obtaining the function, the EM algorithm estimates the parameters with the maximum likelihood of missing data. Then, using these estimations, the best probabilities for the missing data are estimated (Mahmoud & Xia, 2014). In the M-step, new estimates of the parameters are obtained by replacing the estimated missing data and calculating the maximum likelihood over all data (Bruzzone & Prieto, 2002). The EM algorithm iteratively continues the E-step and M-step until the convergence criterion is satisfied (Li, et al., 2019).

3. CLUSTER ANALYSIS OF FOOTBALL PLAYER POSITIONS

Today, the football industry stands out among the sports branches. Especially with the development of technology and its integration into football, different tactical understandings and formations emerge. With these developments, the current positions of the players and the other positions they are prone to play can be revealed as a result of the analysis.

In the study, football players in Turkey (Turkish Super League) and Germany (Bundesliga), England (Premier League), France (Ligue 1), Spain (La Liga), Italy (Serie A) leagues, which are accepted as the five major leagues, according to four basic positions. clustering is intended. In this way, alternative positions where the football players can play alongside their original positions have been put forward based on the current statistical data. Thus, it is thought that it will provide convenience to the clubs and technical committees that want to strengthen their staff. In this direction, the football players discussed were analyzed by using the criteria based on real data on the website called "whoscored". During the analysis, the Weka 3.8 package program, a software developed by the University of Waikato in New Zealand, was used for machine learning.

The football players discussed in the study are divided into four basic positions: goalkeeper, defender, midfielder and attacker. Ancillary sites included in the main sites are given in Table 1. The criteria used during the analysis were determined as the most frequently mentioned attributes in the literature review. The abbreviations of these criteria are given in Table 2. In the literature review, whoscored, transfermarkt and goal websites stand out as providers of player statistics. The "whoscored" website was preferred as the provider that offers the most comprehensive information on player performances. The summary version of the data set used in the study is given in Table 3 and Table 4.

Table 1. Position Table

GOALKEAPER	DEFENSİVE	MIDFIELDER	FORWARD
GK	D	M	AMC
	DL	MC	AM
	DC	ML	AML
	DR	MR	AMR
	DMC		FW

GK: Goalkeaper
D: Defensive
DL: Devensive Left
DC: Defensive Center
DR: Defensive Right
DMC: Defensive Midfielder Center
M: Midfielder
ML: Midfielder Left
MC: Midfielder Center
MR: Midfielder Right
AM: Attacking Midfielder
AML: Attacking Midfielder Left
AMC: Attacking Midfielder Center
AMR: Attacking Midfielder Right
FW: Forward Wings

The football players considered during the analysis were clustered on the basis of 21 criteria in total. These criteria and their abbreviations, obtained from the Whoscored website, are given in Table 2. In the study, the players who took 1500 minutes or more were determined as alternatives; Table 3 and Table 4 give an example of the data set of the study.

Table 2. Criteria and Abbreviations Used During Analysis

CRITERIA	ABBREVIATION	CRITERIA	ABBREVIATION
Goals Per Game	**G.P.G**	Clearance Per Game	**C.P.G**
Assists Per Game	**A.P.G**	Pass Drippled Per Game	**P.D.P.G**
Total Number of Yellow Cards	**T.N.Y.C**	Outfielder Block Per Game	**O.B.P.G**
Total Number of Red Cards	**T.N.R.C**	Key Pass Per Game	**K.P.P.G**
Shooting Average	**S.A**	Dripple Won Per Game	**D.W.P.G**
Pass Percentage	**P.P**	Dispossessed Per Game	**D.P.G**
Aerial Won Per Game	**M.W.P.G**	Total Pass Per Game	**T.P.P.G**
Tackle Per Game	**T.P.G**	Accurate Crosses Per Game	**A.C.P.G**
Interception Per Game	**I.P.G**	Accurate Long Pass Per Game	**A.L.P.P.G**
Fouls Per Game	**F.P.G**	Accurate Through Ball Per Game	**A.T.B.P.G**
Offside Per Game	**O.P.G**		

Table 3. Data Set Used in the Study

G.P.G	A.P.G	T.N.Y.C	T.N.R.C	S.A	P.P	M.W.P.G	T.P.G	I.P.G	F.P.G	O.P.G
4	8	4	0	0,9	92,6	0,9	0,9	0,9	0,9	0
0	5	2	1	0,3	78,4	0	0,9	0,5	0,8	0
4	2	3	1	1,4	80,3	0,9	0,2	0,1	0,7	0
0	1	4	2	0,4	81,6	1,6	0,5	1,3	0,6	0,3
2	6	6	1	0,7	89,2	0,2	0,8	0,1	1,2	0

Table 4. Data Set Used in the Study (Continued)

C.P.G	P.D.P.G	O.B.P.G	K.P.P.G	D.W.P.G	D.P.G	T.P.P.G	A.C.P.G	A.L.P.P.G	A.T.B.P.G
0,6	0,6	0,3	1	0,8	1,2	58,8	0	1,5	0
0,3	0,8	0,1	1,1	1	0,7	20,7	0,6	0,4	0
0,5	0,2	0	0,9	0,6	0,4	22,5	0,1	1,5	0,1
3,6	0,4	0,9	0	0	0	36,3	0	4,7	0
0	0,8	0,1	1,2	2,2	2,9	29,3	0,1	1,5	0,2

The data in Table 3 and Table 4 were adapted to the 'arff' format and loaded into the WEKA package program. Then those that fit the appropriate numerical data set are tried and then the most appropriate one is Expectation Maximization. Since there are four basic positions in football, namely goalkeeper, defense, midfielder and attack, four clusters were determined as the number of clusters in the study. When looking at Figure 1, the result interface that emerges as a result of the analysis using the EM algorithm is seen.

Looking at Figure 1, 91 (9%) of the football players clustered with the EM algorithm were assigned to Cluster 0, 234 (24%) to Cluster 1, 313 (32%) to Cluster 2 and 347 (35%) to Cluster. assigned to 3. In Table 5, the clusters obtained as a result of the analysis made with the EM algorithm and the objects in the clusters are given. The results obtained as a result of the analysis made with the EM algorithm are given in Table 5. Looking at Table 5, it is seen that all of the 91 football players assigned to Cluster 0 belong to the goalkeeper position. Looking at Cluster 1, it was observed that 232 of the 234 football players assigned were players playing in the defensive position. Looking at Cluster 2, it is seen that 219 of the 357 football players assigned to this cluster consist of players playing in the midfield. Finally, when we look at Cluster 3, it is seen that 264 of the 313 appointed players took part in the attacking position. Looking at Table 6, 100% success was achieved in Cluster 1, while 95% success was achieved in Cluster 1, 63% in Cluster 2, and 84% in Cluster 3. In general, 806 of the 985 players were assigned correctly and a success rate of 81.82% was achieved.

Table 5. Result Table Obtained by EM Algorithm

Cluster	Position	Number Of Cluster Data	CLUSTERING RESULT				Percent
			Goalkeaper	Defensive	Midfielder	Attacking	
Cluster 0	**Goalkeaper**	91	**91**	0	0	0	**100,00**
Cluster 1	**Defensive**	234	0	**232**	2	0	**99,15**
Cluster 3	**Midfielder**	347	0	122	**219**	6	**63,11**
Cluster 2	**Attacking**	313	0	14	35	**264**	**84,35**

Table 6. Percentage of Success Obtained with the EM Algorithm

Cluster	Position	Number Of Cluster Data	Right	False	Percent
Cluster 0	**Goalkeaper**	91	91	0	100,00
Cluster 1	**Defensive**	234	232	2	99,15
Cluster 3	**Midfielder**	347	219	128	63,11
Cluster 2	**Attacking**	313	264	49	84,35
TOTAL		985	806	179	**81,82**

4. CONCLUSION AND EVALUATION

Football, which is one of the sports branches that attract the most spectators in the national and international arena, has increased its importance and popularity in recent years. Club managers make transfers to needed positions in addition to the existing staff in order to further increase this visibility. In this study, football players playing in Turkey and in the leagues of England, Italy, Spain, Germany and France, which are considered to be the five major leagues, in the 2020-2021 season are clustered according to four basic positions (goalkeeper, defender, midfielder and attacker). In the analyzes made using the Weka package program, algorithms such as Simple K-Means, Canopy, Expectation Maximization, which are clustering algorithms that can be applied to numerical data suitable for the use of quantitative data, have been tried; It has been seen that the Expectation Maximization EM algorithm gives the most appropriate result. In the evaluation made on the basis of the existing criteria, it was seen that all the players in Cluster 0 were goalkeepers. All 91 players in this cluster are goalkeepers. Looking at Cluster 1, 232 of the 234 players in this cluster are the players in the defensive position. The other two players in this cluster are players who play in the midfield position. According to the results obtained, 264 out of 313 players in Cluster 2 are players playing in the attacking position. While the other 35 players assigned to this cluster are players playing in the midfield, 14 of them are players playing in the defense position. When we look at Cluster 2, the fact that the players who play in the midfield are also included in this cluster is due to the fact that they reveal statistics close to the attacking football players in the data of the 2020-2021 season. Players in the midfield position have created richness for their teams in attack, especially in recent years, with their direct contributions to the score. This made the assignment to Cluster 2 tolerable. Looking at Cluster 3, it is seen that the players playing in the midfield position predominate in this cluster. 219 of the 347 players assigned to the cluster play in the midfield. 122 of the other football players who were appointed consist of players playing in the defense position and 6 of them playing in the attacking position. Considering the other clusters in the study, the highest deviation was experienced in this cluster. One of the reasons for this is the presence of defensive midfielders in football. Being close to each other statistically is one of the reasons why they are in the same cluster.

With the study, the current position of the players was determined and other positions that they can play in addition to their current position were determined. Thus, it will be possible to determine in which positions the current players or the players to be transferred can play based on the data of the past years while making the roster planning of the teams. When there is a choice between more than one player in the transfer planning, the players will be analyzed and will be effective in the decision-making process of the player to be preferred to the desired position. In the training of young players and determining the position they will play, it will contribute to the determination of training programs suitable for the position that emerges with the analysis of the player.

In future studies, by increasing the number of leagues discussed, analysis can be made about the football players in these leagues and different results can be obtained. In addition, in future studies, different leagues can be included in the study, different results can be obtained based on different criteria and by trying different algorithms.

Ethics Committee Approval
N/A

Peer-review
Externally peer-reviewed.

Author Contributions
All authors have read and agreed to the published version of manuscript.

Conflict of Interest
The authors have no conflicts of interest to declare.

Funding
The authors declared that this study has received no financial support.

5. REFERANCES

Akpınar, H. (2014). Data: Veri Madenciliği. İstanbul: Papatya Yayıncılık.

Aygün D, & Ulucenk E. (2019). Futbol Kulüplerinde İnsan Kaynakları Faaliyetlerinin Muhasebeleştirilmesi. *Muhasebe Ve Vergi Uygulamaları Dergisi* , 689–710.

Behravan, I., Razavi, S. M. (2021). A Novel Machine Learning Method For Estimating Football Players' Value İn The Transfer Market. *Soft Computing*, *25*(3), 2499–2511. Https://Doi.Org/10.1007/S00500-020-05319-3

Bruzzone, L., Prieto, F. (2002). An Adaptive Semiparametric and Context-Based Approach to unsupervised Change Detection in Multitemporal Remote-Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 11 (4): 452-466, 2002.

Choudhary, A. (2016). Survey on K-Means and Its Variants. International Journal of Innovative Research in Computer and Communication Engineering, 4(1), ss.949-952.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), ss.1–22.

Gemici, B. (2012). Veri Madenciliği ve Bir Uygulaması. Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı Ekonometri Programı Yüksek Lisans Tezi, İzmir.

Kangalli, S. G. (2014). Oecd Ülkelerinde Ekonomik Özgürlük: Bir Kümeleme Analizi Economic Freedom İn Oecd Countries: A Cluster Analysis. In *Uluslararası Alanya İşletme Fakültesi Dergisi International Journal Of Alanya Faculty Of Business Yıl* (Vol. 6, Issue 3). Http://Www.Heritage.Org/Index/,

Kawasaki, T., Sakaue, K., Matsubara, R., Ishizaki, S. (2019). Football Pass Network Based On The Measurement Of Player Position By Using Network Theory And Clustering. *International Journal Of Performance Analysis İn Sport*, *19*(3), 381–392. Https://Doi.Org/10.1080/24748668.2019.1611292

Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H. (2019). Expectation-MaximizationAttention Networks for Semantic Segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), ss.9167-9176.

Mahmoud, M., Xia, Y. (2014). Expectation Maximization. In Networked Filtering and Fusion in Wireless Sensor Networks.

Narizuka, T., Yamazaki, Y. (2019). Clustering Algorithm For Formations İn Football Games. *Scientific Reports*, *9*(1). Https://Doi.Org/10.1038/S41598-019-48623-1

Öntürk, Y., Karacabey, K., Özbar, N. (2019). Günümüzde Spor Denilince İlk Akla Neden Futbol Gelir? Sorusu Üzerine Bir Araştırma. *Ankara Üniversitesi Beden Eğitimi Ve Spor Yüksekokulu Spormetre Beden Eğitimi Ve Spor Bilimleri Dergisi*, *17*(2), 1–12. Https://Doi.Org/10.33689/Spormetre.533739

Özkan, Y. (2015). Veri Madenciliği Yöntemleri. İstanbul: Papatya Yayıncılık.

Savaş, S., Topaloğlu, N., Yilmaz, M. (2012). Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi Yıl:11 Sayı: 21 Bahar 2012 S. 1-23.*

Stübinger, J., Mangold, B., Knoll, J. (2020). Machine Learning İn Football Betting: Prediction Of Match Results Based On Player Characteristics. *Applied Sciences (Switzerland)*, *10*(1). Https://Doi.Org/10.3390/App10010046

Talimciler, A. (2008). Futbol Değil İş: Endüstriyel Futbol. *İletişim Kuram Ve Araştırma Dergisi Sayı 26 Kış-Bahar*, 89–114.

Tuğbay İ. (2007). Türkiye'deki Futbol Kulüplerinin Gişe Gelirlerini Arttırmaya Yönelik Uygulamaların İncelenmesi. *Çukurova Üniversitesi Sağlık Bilimleri Enstitüsü Beden Eğitimi Ve Spor Anabilim Dalı*.

Yetim, A. A. (2000). Sporun Sosyal Görünümü. *Gazi Beden Eğitimi Ve Spor Bilimleri Dergisi (Gazi Besbd)*, *1*, 63–72.