RICYDE. Revista Internacional de Ciencias del Deporte
ISSN: 1885-3137
ricyde@cafyd.com
Editorial Ramón Cantó Alcaraz
España

Soto-Valero, César

A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system

RICYDE. Revista Internacional de Ciencias del Deporte, vol. XIII, núm. 49, julio, 2017, pp. 244-259
Editorial Ramón Cantó Alcaraz
Madrid, España

RICYDE. Revista Internacional de Ciencias del Deporte
doi:10.5232/ricyde

Rev. int. cienc. deporte

RICYDE. Revista Internacional de Ciencias del Deporte
VOLUMEN XIII - AÑO XIII
Páginas:244-259 ISSN:1885-3137
Número 49 - Julio - 2017

https://doi.org/10.5232/ricyde2017.04904

A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system

Modelo basado en agrupamiento de mixturas Gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos FIFA de EA Sports

César Soto-Valero
Department of Computer Science, Universidad Central "Marta Abreu" de Las Villas, Cuba

Abstract

The generation and availability of football data has increased considerably last decades, mostly due to its popularity and also because of technological advances. Gaussian mixture clustering models represents a novel approach to exploring and analyzing performance data in sports. In this paper, we use principal components analysis in conjunction with a model-based Gaussian clustering method with the purpose of characterizing professional football players. Our model approach is tested using 40 attributes from EA Sports' FIFA video game series system, corresponding to 7705 European players. Clustering results reveal a clear distinction among different performance indicators, representing four different roles in the team. Players were labeled according to these roles and a gradient tree boosting model was used for ranking attributes regarding to its importance. We found that the dribbling skill is the most discriminating variable among the different clustered players' profiles.

Key words: association football; EA Sports' FIFA video game series system; machine learning; principal component analysis; Gaussian mixture clustering models; classification and regression trees.

Resumen

En las últimas décadas se ha visto un incremento considerable en la generación y disponibilidad de datos de fútbol, esto se debe fundamentalmente a la popularidad de este deporte así como a los avances tecnológicos realizados en este campo. Los modelos de agrupamiento basados en mixturas Gaussianas representan un enfoque novedoso para explorar y analizar datos de desempeño deportivo. En el presente trabajo, se lleva a cabo una caracterización de jugadores profesionales de fútbol utilizando técnicas de análisis de componentes principales y agrupamiento basados en mixturas Gaussianas. El modelo presentado es comprobado utilizando datos del sistema de videojuegos FIFA de EA Sports, dichos datos representan 40 atributos correspondientes a 7705 futbolistas europeos. Los resultados del agrupamiento revelan una clara distinción entre algunos indicadores de desempeño, los cuales corresponden a cuatro roles diferentes en el equipo. Consecuentemente, los jugadores fueron etiquetados de acuerdo a estos roles y un modelo de árboles de gradiente ampliado fue utilizado para ordenar los atributos de acuerdo a su importancia. Como resultado se identificó a la habilidad de driblear como la variable que mejor discrimina entre los diferentes perfiles de jugadores.

Palabras clave: fútbol; sistema de videojuegos FIFA de EA Sports; aprendizaje automátic; análisis de componentes principales; agrupamiento basado en modelos de mixturas Gaussianas; árboles de clasificación y regresión.

Correspondencia/correspondence: César Soto-Valero
Department of Computer Science, Universidad Central "Marta Abreu" de Las Villas, Cuba
Email: cesarsotovalero@gmail.com

Received: January 9, 2017; Accepted: March 29, 2017

Introduction

Association football, also known as soccer, is recognized for being a globally played sport; its popularity has increased significantly in the last decades. In particular, the Union of European Football Associations (UEFA) represents the most important confederation, which is directly supported by the International Federation of Association Football (FIFA). In this context, the European clubs have transformed into business organizations (Moor, 2007), making professional football a multi-billion dollar business (Lanfranchi & Taylor, 2001).

Technological advances have increased the generation and availability of quantitative football data. For example, modern video analysis (Shu-Ching, Mei-Ling, & Na, 2005) and sophisticated notation systems (James, 2006) allow to register and store a huge amount of specific information about players' actions during every match played (i.e. ball controlling, passing, shoots, etc.). This has facilitated to define and identify a large number of critical elements of players' performance (M. Hughes & Franks, 2005). The EA Sports' FIFA video game series represents a valuable effort, offering detailed information about players and team my means of a complete set of quantitative attributes (Markovits & Green, 2017).

In football, each player is assigned to one of the 11 particular positions on the field of play. These positions represent both the player's main role and their area of operation on the pitch. The problem of characterizing football players according to their position on the field is complex because of the fluid nature of the modern game. Players' positions are not rigidly defined, increasing the number of "utility players" who are able to play comfortably in various roles. Even so, most players will play in a limited range of positions throughout their career, as each position requires a particular set of skills and physical attributes.

Players' scouting and training design are both important coaching skills. Coaches rely on various heuristics (e.g., wage, special abilities, own experience and intuition, etc.) to select a specific football player for their teams (Bidaurrazaga Letona, Lekue, Amado, Concejero, & Gil, 2015). However, with the rapid increase in the volume of football data in digital form, the use of specific metrics for characterizing and ranking players according to their perceived abilities has attracted the attention of coaches and data scientists. The use of tools for analyzing the performance of professional football players based on available data could represent an important competitive advantage. This field of research has received an increased support nowadays by the football leading community (McCall et al., 2016; Sarmento et al., 2014).

In this scenario, it is clear the necessity of tools that can effectively search for interesting information in large football datasets. Data mining is a field of computer science which deals with discovering interesting patterns in data. An important step in data mining process is the application of machine learning methods, which is related to obtaining and managing knowledge from data. Machine learning methods have been successfully applied in football. For example, in the prediction of match outcomes (Constantinou, Fenton, & Neil, 2012; Min, Kim, Choe, Eom, & McKay, 2008; Odachowski & Grekow, 2013; Strnad, Nerat, & Kohek, 2015; Tüfekci, 2016), analysis of team performance (Arruda Moura, Barreto Martins, & Augusto Cunha, 2013) or player's injury prediction (Arndt & Brefeld, 2016; Jelinek, Kelarev, Robinson, Stranieri, & Cornforth, 2014; Kampakis, 2011). However, the problem of characterizing and selecting players based on available data of performance using machine learning methods is an interesting and open field of research today.

Cluster analysis is one of the main tasks of machine learning; its purpose is to organize unlabeled data into different groups according to some defined similarity rules (Xu & Wunsch, 2009). Finite mixture models represent an useful approach to model a wide variety of random phenomena for clustering, classification and density estimation (Scrucca, Fop, Murphy, & Raftery, 2016). Thus, it is possible to model football players' data as a Gaussian finite mixture, with different covariance structures and different numbers of mixture components, for a variety of purposes of analysis.

The main contribution of this paper consists of presenting a model-based clustering method for selecting professional football players using multivariate data of performance. The validation dataset consists in a variety of statistics from European professional football gathered by the EA Sports' FIFA video game series system. Firstly, we use principal component analysis in order to reduce data dimensions and then Gaussian finite mixture clustering will by applied for grouping the data. This procedure selects players according to their clustered characteristics and then labels them according to different roles. Our purpose consists in discovering the common characteristics of players related to each role, as well as the profiles of border line players (those with no clear distinction between groups).

Methods

Principal components analysis

One of the most challenging aspects of multivariate data analysis is the reduction of its dimensionality. A statistical technique for exploring and simplifying complex multivariate data is known as principal components analysis (PCA). The goal of PCA is to replace a large number of possible correlated variables with a much smaller set of uncorrelated variables, while capturing as much information in the original variables as possible (Kabacoff, 2011). These derived variables, called principal components, are linear combinations of the observed variables. Specifically, the first principal component (Equation 1) is the weighted combination of the k observed variables that accounts for the most variance in the original set of variables.

$$PC_1 = a_1X_1 + a_2X_2 + \ldots + a_kX_k \quad \textbf{(1)}$$

The second principal component is the linear combination that accounts for the most variance in the original variables, under the constraint that it's orthogonal (uncorrelated) to the first principal component (Jolliffe, 2002). Each subsequent component maximizes the variance accounted for, while at the same time remaining uncorrelated with all previous components. It is possible to extract as many principal components as there are variables. However, from a practical viewpoint the common task is to approximate the full set of variables with a much smaller set of components. This procedure has the advantage that allows data visualization.

PCA has been used extensively in sports data analysis including football. For example, Barros, Cunha, Magalhães, and Guimarães (2006) applied PCA to represent and quantify the pitch region used by different football players and, using these analyses, to provide tactical information about the team; and Arruda Moura et al. (2013) explored football game-related statistics during football competitions in order to group and distinguish variables related to different game outcomes.

Gaussian finite mixture model-based clustering

Cluster analysis identifies groups of observations that are cohesive and separated from other groups, interest in clustering data has experienced a recent surge in sport sciences (Andrienko,

Andrienko, Budziak, von Landesberger, & Weber, 2016; Filipcic, Panjan, & Sarabon, 2014). Among the different clustering methods available, those based on probability models rather than heuristic procedures are becoming increasingly common due to recent advances in methods and software for model-based clustering, and the fact that the results are more easily interpretable. Finite mixture models (McLachlan & Peel, 2004), in which each component probability corresponds to a cluster, provide a principled statistical approach to clustering. Thus, models that differ in the number of components and component distributions can be compared using statistical criteria. The clustering process estimates a model for the data that allows for overlapping clusters, as well as a probabilistic clustering that quantifies the uncertainty of observations belonging to components of the mixture.

Let $\mathrm{x} = \{\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}\}$ be a sample of n independent identically distributed observations, in model-based clustering these are viewed as coming from a mixture density $\mathrm{f(x)} = \sum_{\mathrm{k}=1}^{\mathrm{G}} \tau_\mathrm{k} \mathrm{f}_\mathrm{k}(\mathrm{x})$, where f_k is the probability density function of the observations in group k, and τ_k is the probability that an observation comes from the kth mixture component $(0 < \tau_\mathrm{k} < 1)$, for all $\mathrm{k} = \{1, \ldots, \mathrm{G}\}$ and $\sum_\mathrm{k} \tau_\mathrm{k} = 1$. Each component is usually modeled by the normal or Gaussian distribution. In the multivariate case, component distributions are characterized by the mean μ_k and the covariance matrix $\sum_\mathrm{k}$, and the probability density function is expressed as show in Equation 2.

$$\phi\left(\mathrm{x_i};\mu_\mathrm{k}, \Sigma_\mathrm{k}\right) = \frac{\exp\{-{}^1/_2(x_i-\mu_\mathrm{k})^T\, {}^1/_{\Sigma_\mathrm{k}}\, (x_i-\mu_\mathrm{k})\}}{\sqrt{\det(2\pi \Sigma_\mathrm{k})}} \quad \textbf{(2)}$$

The likelihood for data consisting of n observations, assuming a Gaussian mixture model with G multivariate mixture components, is $\prod_{\mathrm{i}=1}^{\mathrm{n}} \sum_{\mathrm{k}=1}^{\mathrm{G}} \tau_\mathrm{k}\phi(\mathrm{x_i};\mu_\mathrm{k}, \Sigma_\mathrm{k})$. For a fixed number of components G, the model parameters τ_k, μ_k, and $\sum_\mathrm{k}$ can be estimated via the EM algorithm (Dempster, Laird, & Rubin, 1997) and initialized by hierarchical model-based clustering (Dasgupta & Raftery, 1998). Data generated by mixtures of multivariate normal densities are characterized by groups or clusters centered at the means μ_k, with increased density for points nearer the mean. The corresponding surfaces of constant density are ellipsoidal.

Geometric features (shape, volume, orientation) of the clusters are determined by the covariances $\sum_\mathrm{k}$, which may also be parameterized to impose cross-cluster constraints. There are a number of possible parameterizations of $\sum_\mathrm{k}$, Banfield and Raftery (1993) proposed a general framework for geometric cross-cluster constraints in multivariate normal mixtures by parameterizing covariance matrices through eigenvalue decomposition in the form of Equation 3.

$$\Sigma_\mathrm{k} = \lambda_\mathrm{k} \mathrm{D}_\mathrm{k} \mathrm{A}_\mathrm{k} \mathrm{D}_\mathrm{k}^\mathrm{T} \quad \textbf{(3)}$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues, and λ_k is an associated constant of proportionality. The idea is to treat λ_k, D_k, and A_k as independent sets of parameters, and either constrain them to be the same for each cluster or allow them to vary among clusters. When parameters are fixed, clusters will share certain geometric properties: D_k governs the orientation of the kth component of the mixture, A_k its shape, and λ_k its volume, which is proportional to $\lambda_\mathrm{k}^\mathrm{d} \det(\mathrm{A}_\mathrm{k})$. Therefore, 14 possible models with different geometric characteristics can be specified (Scrucca et al., 2016).

A "best" clustering model for the data can be estimated by fitting models with differing parameterizations or numbers of clusters to the data by maximum likelihood, and then applying a statistical criterion for model selection (Fraley & Raftery, 1998). The Bayesian Information Criterion or BIC (Schwartz, 1978) is the model selection criterion most commonly used; the "best" model is taken to be the one with the highest BIC value.

Classification and regression trees

In machine learning and data mining, classification and regression trees (CART) are a non-parametric group of methods based on the construction of decision trees for induction. CART are a bit different from decision trees, where the leaf only contains decision values. In CART, a real score is associated with each of the leaves, which gives a richer interpretation of the problem that goes beyond classification. This technique produces either classification or regression trees, depending on whether the dependent attribute is categorical or numeric, respectively.

In order to construct a decision tree, CART algorithms usually work top-down, by separating recursively instances into branches with the purpose of achieving the highest possible prediction accuracy. In doing so, different mathematical metrics (e.g., Chi-square statistics, Information Gain, Gini Index, etc.) could be used to identify an attribute and its corresponding threshold, with the purpose of splitting the instances into two or more subgroups (Han & Kamber, 2006). The model splitting continues at each leaf node until the model's explanatory power (on a training dataset) is not further improved by additional splits (or the created subsets are too small to be subdivided).

Nowadays, one of the most popular CART algorithms is gradient tree boosting (Friedman, 2001). The algorithm is a fast tree ensemble model that built a scalable set of classification and regression trees. It uses a sparsity-aware technique for sparse data and a weighted quantile sketch procedure that enables handling instance weights for approximate tree learning (Chen & Guestrin, 2016). The algorithm has been used effectively in a wide range of machine learning and data mining problems, finding application in areas such as medicine, financial analysis and astronomy.

Dataset

Since 2009, the EA Sports' FIFA video game series system offers detailed information, including weekly updates, about a large set of European football players and teams attributes. This data is available for free from its official website (http://sofifa.com/). FIFA series and all FIFA assets are property of EA Sports. The system has resulted in a huge amount of fine-grained data; which has proven to be especially useful for coaches, sports analysts and fans of football worldwide (Markovits & Green, 2017).

Recently, Mathien (2016) compiled, cleaned and shared a dataset of statistics of the European professional football. He used the EA Sports' FIFA video game series system for organizing a SQL database which includes a characterization of more than 10000 players from the top football leagues in 11 European countries. This impressive collection of data allows finding insights about the footballers' performance onto a quantitative perspective.

In this work, we used the players' statistics of the Mathien's database but also we extended it with general characteristics such as age and BMI (Bloomfield, Polman, Butterly, & O'Donoghue, 2005). To have more robust stats, we removed players whose score is available for less than 10 matches and then compute average scores for each player with more than 10 matches played. Thus, the total number of selected players for our analysis was 7705. Table 1 shows the mean and standard deviations (SD) values of the 40 attributes used in our analysis (excluding players' names), dated on October 2016.

Table 1: Mean and standard deviations of players' attributes according to EA Sports FIFA games system (October 2016).

Type	Attribute	Mean	SD	Type	Attribute	Mean	SD
Physical	Age	25.68	4.12	Defensive	Marking	46.63	20.22
	Height	181.9	6.38		Standing Tackle	50.08	20.49
	Weight	168.9	15.12		Sliding Tackle	47.78	20.76
	BMI	23.11	1.34				
Attacking	Crossing	54.32	16.22	Mentality	Aggression	60.59	14.65
	Finishing	48.95	18.16		Interceptions	51.63	17.41
	Heading Accuracy	56.78	15.78		Positioning	54.71	16.72
	Short Passing	61.74	13.32		Vision	57.00	13.93
	Volleys	48.28	17.55		Penalties	54.09	13.88
Movement	Acceleration	67.10	11.87	Goalkeeper	Diving	15.13	17.13
	Sprint Speed	67.51	11.38		Handling	16.34	15.80
	Agility	65.41	12.12		Kicking	20.75	15.36
	Reactions	65.61	7.48		Positioning	16.42	16.05
	Balance	64.81	11.23		Reflexes	16.76	17.18
Skill	Dribbling	58.24	17.01	Power	Shot Power	60.86	15.15
	Curve	51.93	17.46		Jumping	66.75	9.50
	Free Kick Accuracy	48.43	16.55		Stamina	66.46	11.31
	Long Passing	56.45	12.70		Strength	67.30	10.78
	Ball Control	62.61	14.54		Long Shots	52.24	17.31
General	Potential	73.06	5.41	General	Overall Rating	68.19	5.65

Procedures

The objective of this paper is to propose a model-based cluster method which enables characterizing professional football players according to their role and performance criteria. The layout of the method, which is tested in our experiments, is illustrated in Figure 1.

All the experiments were developed using the R statistical computing software (version 3.3.2). First, we performed a principal components analysis in order to summarize and display graphically attributes from raw data, for this aim we used the ggplot2 R package (Wickham, 2015). Players' roles were obtained via Gaussian mixture clustering, using the mclust R package, from these principal components (Scrucca et al., 2016). Next, the original data was labeled according to the previously defined players' roles and an extreme gradient tree boosting classifier was trained to perform feature selection (ranking) and classification analysis using the xgboost R package (Chen & Guestrin, 2016).

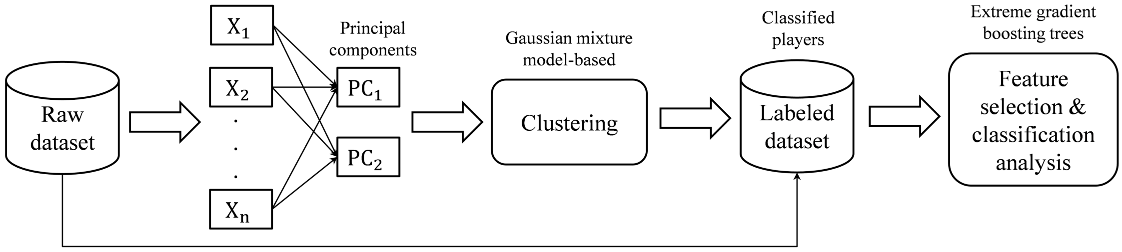


Figure 1: Graphical representation of the methodology followed for analyzing our football players' dataset.

Results

Figure 2 shows the two first principal components (PC_1 and PC_2) obtained, which represent linear combinations of the 40 raw variables of data. These components were selected based on the eigenvalues criterion to maximize the variance, while keeping the components uncorrelated. The first PC is associated with the largest eigenvalue (accounts for 45.2 percent of the variance), and the second PC with the second largest eigenvalue (15.9 percent of the variance).

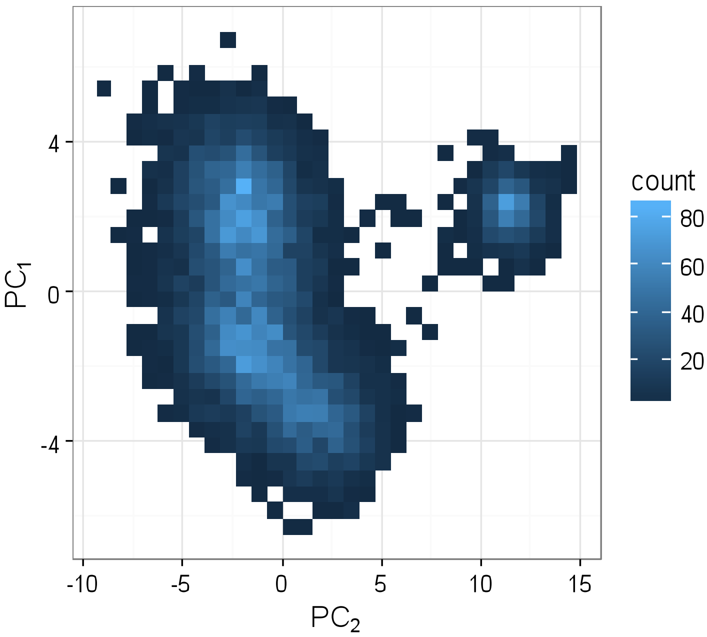


Figure 2: Scatter plot projection of the first two principal components.

The PCA graph revealed the presence of two major clusters. To precisely define these clusters, we used a Gaussian mixture model-based clustering. Figure 3 shows the BIC traces during the model fitting procedure by maximum likelihood. In this case, the best (with the highest BIC value) was the ellipsoidal model VEV, adjusted with a total of four components (clusters). For more details about this fitting model see Browne and McNicholas (2014) and Scrucca et al. (2016).

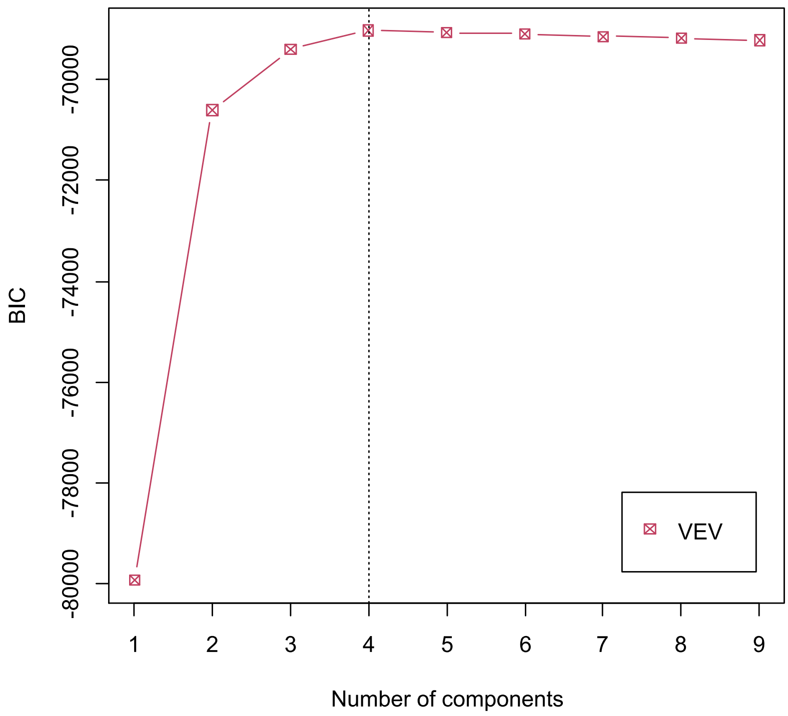


Figure 3: BIC plot for the VEV model fitted to the football players' dataset. The dotted line indicates the best fit.

Figure 4 represents a projection of the clusters obtained by the model, with different colors (red, green, blue and purple) indicating the classification for 3062, 1457, 2511 and 675 players respectively. The four component means, selected by the method, are marked and ellipses with axes are drawn corresponding to their covariance.

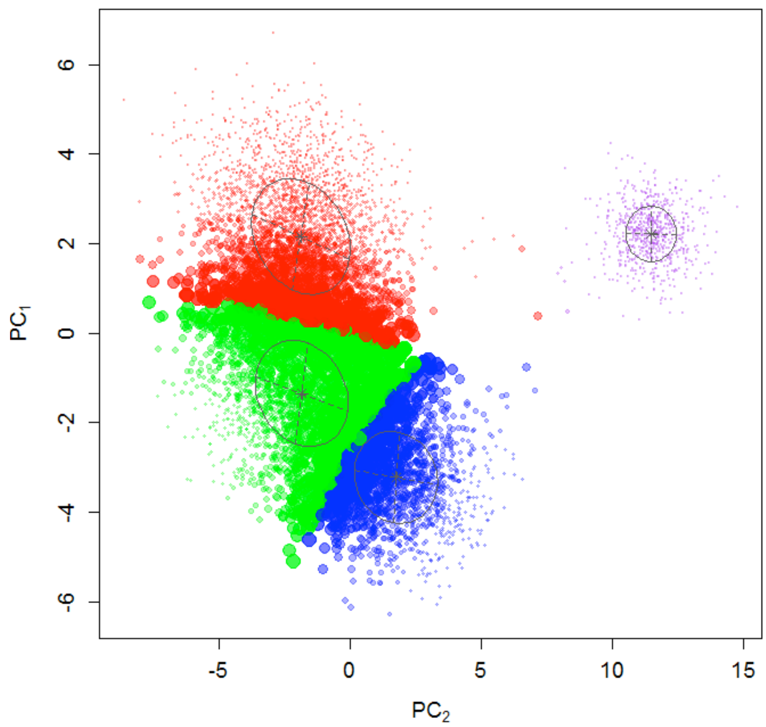


Figure 4: Projection of the players' dataset showing classification clustering uncertainty. Larger points indicate the more uncertain observations.

To make sense of the four clusters obtained we made a deep look at which players they contain. Thus, we found that the four clusters correspond to different roles in the team: defenders (blue), midfielders (green) forwards (red) and goalkeepers (purple). Figure 5 compares these profiles using a radar chart of mean values by attribute. It is clear that some variables related directly to goalkeepers (Diving, Handling, Kicking, Positioning and Reflexes) have lower values to the rest of profiles. On the other hand, general attributes (Overall Rating and Potential) show no significant differences in its mean values of performance among all players' roles.

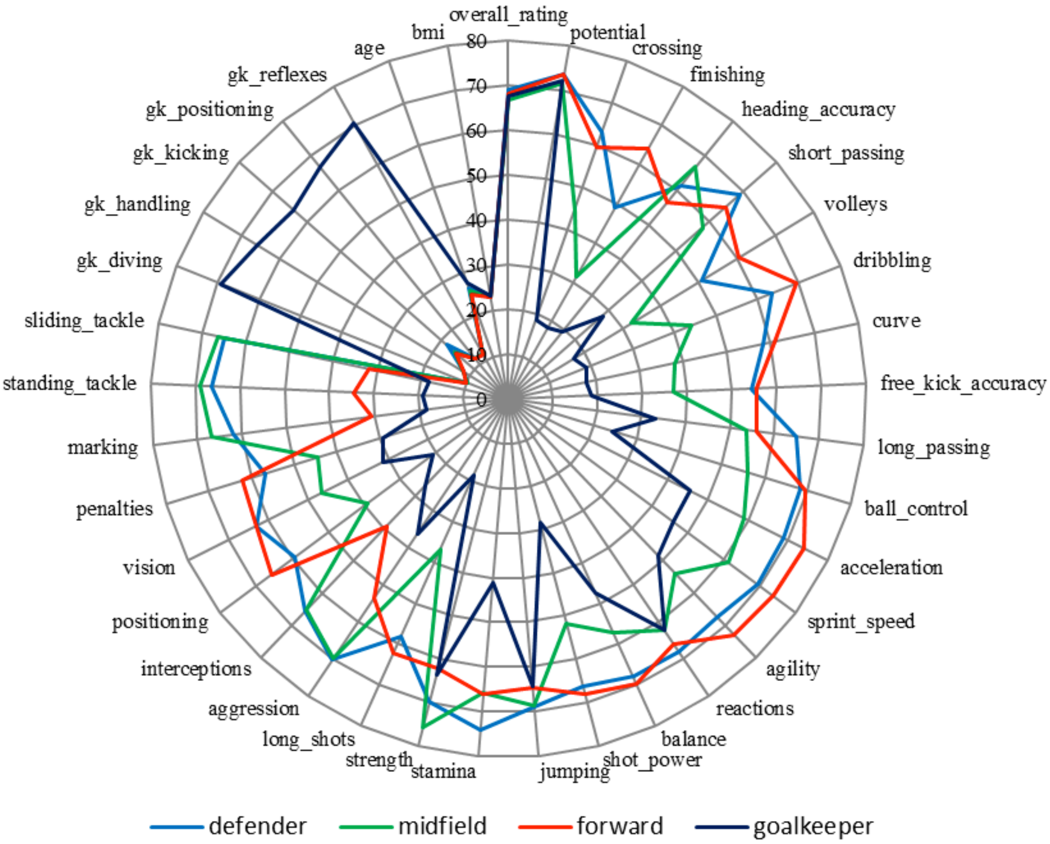
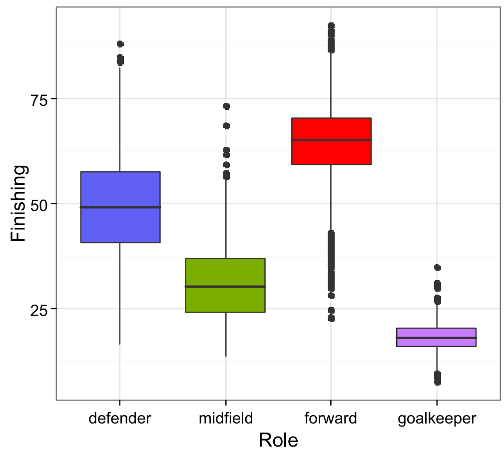
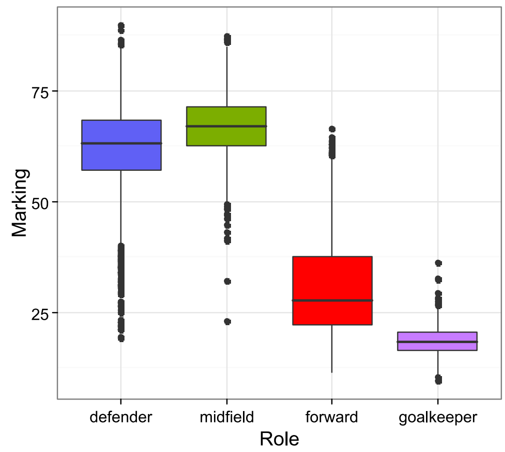


Figure 5: Radar chart of mean values for each attribute grouped by players' roles.

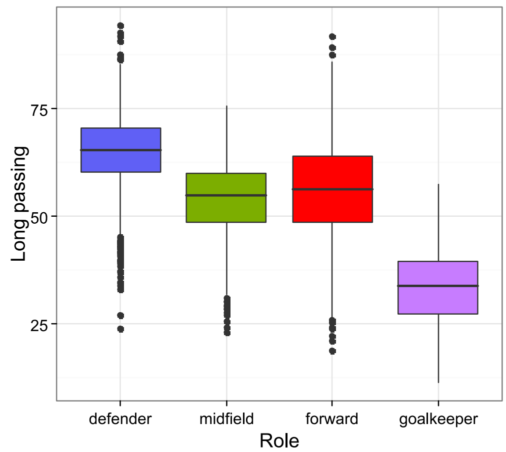
We want to explore the distribution of some attributes among players in a little more detail. Figure 6 shows box plots representing a comparison of four different attributes. According to these figures, it is easy to note that some types of attributes are directly related with different players' roles. For example, attacking and defense attributes correlates well with forward and defender players.



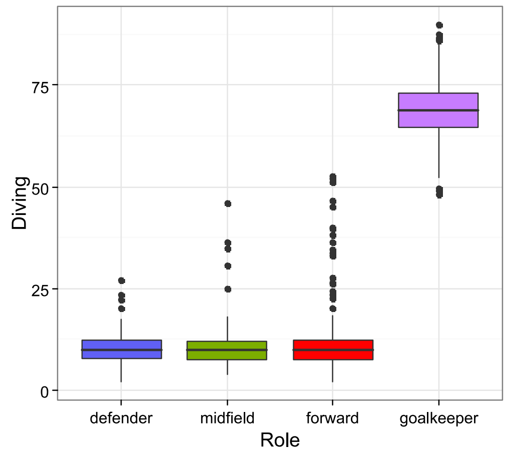
(a)



(b)



(c)



(d)

Figure 6: Box plots of players' attributes that correlates well to different roles; (a) finishing with forward, (b) marking with midfield, (c) long passing with defender and (d) diving with goalkeeper.

Following the methodology showed in Figure 1, our next step consists in labeling the initial dataset according to the clustering results. Consequently, we assigned their corresponding classification role obtained from clustering to each football player. This procedure enables us to apply machine learning methods for feature selection and classification analysis (Han & Kamber, 2006).

We fit a gradient tree boosting model using the cluster labeled dataset. The model uses linear regression as its objective function. The overall performance of the model was evaluated using stratified 10-fold cross-validation (Witten, Frank, & Hall, 2011). The values of accuracy, sensitivity and specificity obtained were 0.95, 0.96 and 0.98 respectively. Thus, in general, we can claim that the model performs properly.

We used the fitted model in order to evaluate the relative importance of each attribute. With this method, each split in the decision tree tries to find the best feature and splitting point to optimize the classification objective (Morgan, Williams, & Barnes, 2013). Thus, the Gain value can be calculated on each node, and it is the contribution from the selected attribute. Gain is the improvement in accuracy brought by an attribute to the branches it is on.

In the end, we look into all the trees, and sum up all the contribution for each feature and treat it as the importance. Figure 7 shows the attribute importance plot resulted from this procedure. The size of each horizontal bar represents the relative importance of each attribute with respect to the rest of the predictive variables.

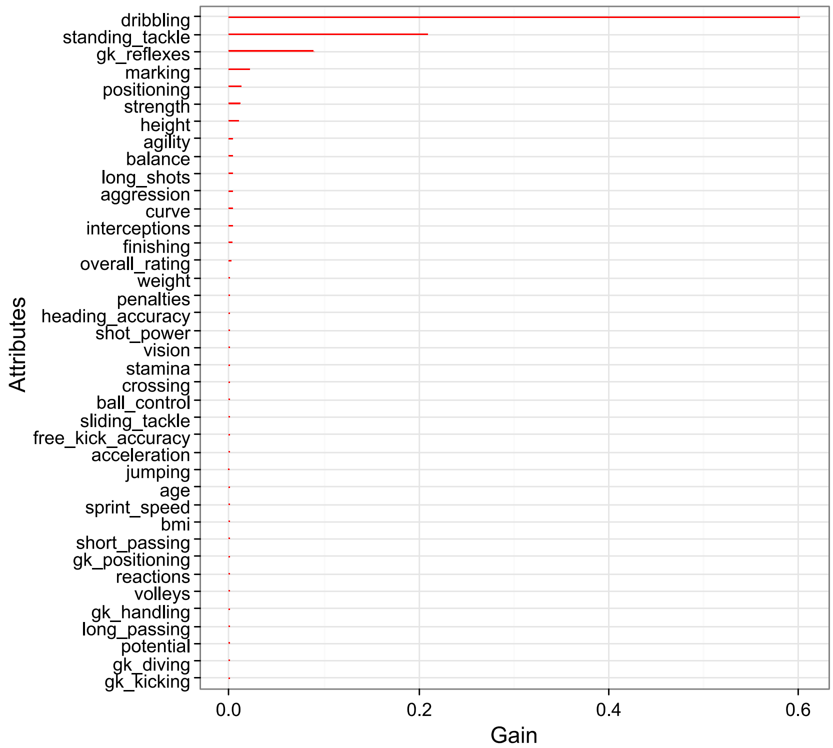


Figure 7: Horizontal bar chart representing the attributes' importance ranked according to their Gain values.

According to the plot above, the most important attribute to differentiate players according to their role is the dribbling skill (Gain = 0.6). Also, we found that standing tackle (Gain = 0.21) and reflexes (Gain = 0.9) are both good attributes to characterize defenders and goalkeepers respectively.

Discussion

This paper has introduced a model for classifying professional football players using multivariate data from the EA Sports' FIFA video game series system. The proposed method handles the issue by clustering a set of players' attributes, which represent general measured variables of performance (M. D. Hughes & Bartlett, 2002).

Raw data was first summarized using PCA and then each player was labeled according to his clustering classification. Results show a clear distinction between attributes regarding to goalkeepers and the other positions, which is in line with previous studies about characteristics of football players (Di Salvo et al., 2007; Erkmen, 2009; Gil, Gil, Ruiz, Irazusta, & Irazusta, 2007). The clustering method applied also identifies four main groups of players according to its BIC values (Figure 3). These groups represent association football main positions, showing that the demands on the physical and technical realms are different depending on the specific position the player takes in the field (Sarmento et al., 2014).

The entire dataset was then labeled according to the clustering results. The new dataset was used for performing an attribute ranking procedure using a gradient tree boosting model for feature selection. We found that dribbling (the act of taking the ball forward passing opponents with slight touches of the feet) results the most discriminant attributes among the four different players' roles identified (Figure 7). This reveals the importance of dribbling, which is recognized as one of the most difficult ball skills to master and one of the most useful attacking moves in association football.

We found some outliers when plotting players' attributes (Figure 6). Accordingly, we note that some defensive and offensive players still have relatively high average scores in activities associated to goal keepers, such as diving or handling. This may be due to misclassification during the clustering process, but also give us some doubts on the reliability of these scores.

On the other hand, our results were in agreement with previous studies, where authors hypothesized that there are significant differences in football players related to their physical characteristics and playing position (Pau, Ibba, Leban, & Scorcu, 2014; Romann & Fuchslocher, 2013). In this sense, players with good records of positioning and sprint speed tend to be, on average, classified as forwards, while other skills such as long and short passing are more specific of defenders. The search for reliable variables of performance is an important field for talent identification and development in football (Reilly, Williams, Nevill, & Franks, 2000).

One interesting advantage provided by Gaussian mixture clustering models is that it makes possible to find players with any desired proportion of these characteristics (e.g., midfield vs. forward). For example, with this method a coach can select players with well-defined roles, which translates to very high probabilities to belong to the group of defenders, midfielders, forwards or goalkeepers. By playing with the balance defensive/offensive skills, a coach would have an objective criterion to select the players. Also, could be interesting to find for "utility" players, who turn out to be, for example, 50 percent characterized as forward and 50 percent as midfield.

The proposed methodology on this paper can be extended in order to describe football teams and also for performing more general tactical analyses (Memmert, Lemmink, & Sampaio, 2017). Teams could be clustered according to previously well-defined indicators of performance (Cintia, Giannotti, Pappalardo, Pedreschi, & Malvaldi, 2015), which allow to find characteristics associated to similar football teams, for example, with they play style (Gyarmati, Kwak, & Rodriguez, 2014) or even to predict and explain match outcomes (Spencer, Morgan, Zeleznikow, & Robertson, 2016). There is a large potential in machine learning methods to boost the understanding of the football game. This is an important area in quantitative analysis of sports data that requires further research.

Summarizing, players' scouts and training designers need to use modern tools for characterizing and ranking players. In this context, quantitative analysis of multivariate data of performance using machine learning methods, such as clustering or classification, represents an important step into this process. Overall, this study provides further insight concerning player's characterization using freely available data of performance. However, other components such as cognitive and psychological factors must be taking into account due to its proven importance to excel in football.

Conclusion

This paper proposes a novel model-based clustering method for selecting and ranking professional football players according to multivariate data of performance. In addition, it is proposed a framework for grouping and selecting player using free available data. The model uses Gaussian finite mixtures in order to group and classify players according to performance attributes using probabilistic and statistical criteria. In order to test our model with real data, we used statistics from 7705 European professional football players gathered by the EA Sports' FIFA video game series system. The results show a clear role distinction associated with each cluster, corresponding to the well-known main football positions: defender, midfield, forward and goalkeeper. The attribute ranking method, based on gradient boosted trees, determined that the most discriminant variable among the different players' profiles is the dribbling skill. This work supports the conception that each position of players in football is defined by specific performance indicators. The application of this model to an extended set of variables and real data could reveal more insights about the characteristics of some specific football players, which is inestimable for a correct selection of players during scouting.

References

Andrienko, G.; Andrienko, N.; Budziak, G.; von Landesberger, T., & Weber, H. (2016). Coordinate Transformations for Characterization and Cluster Analysis of Spatial Configurations in Football. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part III* (pp. 27–31). Cham: Springer International Publishing.

Arndt, C., & Brefeld, U. (2016). Predicting the Future Performance of Soccer Players. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 9*(5), 373–382. http://dx.doi.org/10.1002/sam.11321

Arruda Moura, F.; Barreto Martins, L. E., & Augusto Cunha, S. (2013). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences.* http://dx.doi.org/10.1080/02640414.2013.853130

Banfield, J., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*(3), 803–821. http://dx.doi.org/10.2307/2532201

Barros, R. M. L.; Cunha, S. A.; Magalhães, J. W. J., & Guimarães, M. F. (2006). Representation and analysis of soccer players' actions using principal components. *Journal of Human Movement Studies, 51*, 103–116.

Bidaurrazaga Letona, I.; Lekue, J. A.; Amado, M.; Concejero, J. S., & Gil, S. M. (2015). Identifying talented young soccer players: conditional, anthropometrical and physiological characteristics as predictors of performance. *RICYDE. Revista internacional de ciencias del deporte, 33*(11), 75-95. http://dx.doi.org/10.5232/ricyde2015.03906

Bloomfield, J.; Polman, R.; Butterly, R., & O'Donoghue, P. (2005). Analysis of age, stature, body mass, BMI and quality of elite soccer players from 4 European leagues. *The Journal of Sports Medicine and Physical Fitness, 45*(1), 58-67.

Browne, R. P., & McNicholas, P. D. (2014). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification, 8*(2), 217–226.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Paper presented at the Proceedings of the 22Nd ACM SIGKDD Internacional Conference on Knowledge Discovery and Data Mining.

Cintia, P.; Giannotti, F.; Pappalardo, L.; Pedreschi, D., & Malvaldi, M. (2015, 19-21 Oct. 2015). *The harsh rule of the goals: Data-driven performance indicators for football teams.* Paper presented at the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Constantinou, A. C.; Fenton, N. E., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems, 36*, 322-339. http://dx.doi.org/10.1016/j.knosys.2012.07.008

Dasgupta, A., & Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via modelbased clustering. *Journal of the American Statistical Association, 93*, 294–302.

Dempster, A. P.; Laird, N. M., & Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39*(1), 1–38.

Di Salvo, V.; Baron, R.; Tschan, H.; Calderon Montero, F. J.; Bachl, N., & Pigozzi, F. (2007). Performance characteristics according to playing position in elite soccer. *International Journal of Sports Medicine, 28*(3), 222–227. http://dx.doi.org/10.1055/s-2006-924294

Erkmen, N. (2009). Evaluating the heading in profesional soccer players by playing position. *Journal of Strength and Conditioning Research, 23*(6), 1723-1728. http://dx.doi.org/10.1519/JSC.0b013e3181b42633

Filipcic, A.; Panjan, A., & Sarabon, N. (2014). Classification of top male tennis players. *International Journal of Computer Science in Sport, 13*(1),

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578-588. http://dx.doi.org/10.1093/comjnl/41.8.578

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29*(5), 1189-1232.

Gil, S. M.; Gil, J.; Ruiz, F.; Irazusta, A., & Irazusta, J. (2007). Physiological and anthropometric characteristics of young soccer players according to their playing position: relevance for the selection process. *The Journal of Strength & Conditioning Research, 21*(2), 438-445.

Gyarmati, L.; Kwak, H., & Rodriguez, P. (2014). Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308.*

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques* (2nd ed.): Morgan Kaufmann Publishers.

Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences, 23*(5), 509–514. http://dx.doi.org/10.1080/02640410410001716779

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of Sport Sciences, 20*(10), 739-754. http://dx.doi.org/10.1080/026404102320675602

James, N. (2006). Notational analysis in soccer: Past, present and future. *International Journal of Performance Analysis in Sport, 6*(2), 67–81.

Jelinek, H. F.; Kelarev, A.; Robinson, D. J.; Stranieri, A., & Cornforth, D. J. (2014). Using meta-regression data mining to improve predictions of performance based on heart rate dynamics for Australian football. *Applied Soft Computing, 14*, 81-87. http://dx.doi.org/10.1016/j.asoc.2013.08.010

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). NY: Wiley Online Library.

Kabacoff, R. I. (2011). Principal components and factor analysis. In *R in Action*. Shelter Island, NY: Manning Publications Co.

Kampakis, S. (2011). *Comparison of machine learning methods for predicting the recovery time of professional football players after an undiagnosed injury.* Paper presented at the Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2013 workshop, Prague, Czech Republic.

Lanfranchi, P., & Taylor, M. (2001). *Moving with the ball: the migration of professional footballers*. Oxford: Berg.

Markovits, A. S., & Green, A. I. (2017). FIFA, the video game: a major vehicle for soccer's popularization in the United States. *Sport in Society, 20*(5-6), 716-734 http://dx.doi.org/10.1080/17430437.2016.1158473

Mathien, H. (2016). Football data collection. From: https://github.com/hugomathien/football-data-collection

McCall, A.; Davison, M.; Carling, C.; Buckthorpe, M.; Coutts, A. J., & Dupont, G. (2016). Can off-field "brains" provide a competitive advantage in professional football? *Journal of Sports Medicine, 50*, 710-712.

McLachlan, G., & Peel, D. (2004). *Finite Mixture Models*: John Wiley & Sons.

Memmert, D.; Lemmink, K. A. P. M., & Sampaio, J. (2017). Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. *Sports Medicine, 47*(1), 1-10.
http://dx.doi.org/10.1007/s40279-016-0562-5

Min, B.; Kim, J.; Choe, C.; Eom, H., & McKay, R. B. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems, 21*(7), 551-562.
http://dx.doi.org/10.1016/j.knosys.2008.03.016

Moor, L. (2007). Sport and commodification: A reflection on key concepts. *Journal of Sport and Social Issues, 31*(2), 128-142.

Morgan, S.; Williams, M. D., & Barnes, C. (2013). Applying decision tree induction for identification of important attributes in one-versus-one player interactions: A hockey exemplar. *Journal of Sports Sciences, 31*(10), 1031-1037.
http://dx.doi.org/10.1080/02640414.2013.770906

Odachowski, K., & Grekow, J. (2013). Using Bookmaker Odds to Predict the Final Result of Football Matches. In M. Graña, C. Toro, R. J. Howlett & L. C. Jain (Eds.), *Knowledge Engineering, Machine Learning and Lattice Computing with Applications: 16th International Conference, KES 2012, San Sebastian, Spain, September 10-12, 2012, Revised Selected Papers* (pp. 196-205). Berlin, Heidelberg: Springer Berlin Heidelberg.

Pau, M.; Ibba, G.; Leban, B., & Scorcu, M. (2014). Characterization of Static Balance Abilities in Elite Soccer Players by Playing Position and Age. *Research in Sports Medicine, 22*(4), 355-367.
http://dx.doi.org/10.1080/15438627.2014.944302

Reilly, T.; Williams, A. M.; Nevill, A., & Franks, A. (2000). A multidisciplinary approach to talent identification in soccer. *Journal of sports sciences, 18*(9), 695-702. http://dx.doi.org/10.1080/02640410050120078

Romann, M., & Fuchslocher, J. (2013). Influences of player nationality, playing position, and height on relative age effects at women's under-17 FIFA World Cup. *Journal of Sports Sciences, 31*(1), 32-40. http://dx.doi.org/10.1080/02640414.2012.718442

Sarmento, H.; Marcelino, R.; Anguera, M. T.; CampaniÇo, J.; Matos, N., & LeitÃo, J. C. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences, 22*(20), 1831–1843. http://dx.doi.org/10.1080/02640414.2014.898852

Schwartz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464. http://dx.doi.org/10.1214/aos/1176344136

Scrucca, L.; Fop, M.; Murphy, B. T., & Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal, 8*(1), 289-317.

Shu-Ching, C.; Mei-Ling, S., & Na, Z. (2005). *An enhanced query model for soccer video retrieval using temporal relationships.* Paper presented at the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan.

Spencer, B.; Morgan, S.; Zeleznikow, J., & Robertson, S. (2016). *Clustering team profiles in the Australian Football League using performance indicators.* Paper presented at the Proceedings of the 13th Australasian Conference on Mathematics and Computers in Sport, Melbourne.

Strnad, D.; Nerat, A., & Kohek, S. (2015). Neural network models for group behavior prediction: a case of soccer match attendance. *Neural Computing and Applications*, 1-14. http://dx.doi.org/10.1007/s00521-015-2056-z

Tüfekci, P. (2016). Prediction of Football Match Results in Turkish Super League Games. In A. Abraham, K. Wegrzyn-Wolska, E. A. Hassanien, V. Snasel & M. A. Alimi (Eds.), *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015* (pp. 515-526). Cham: Springer International Publishing.

Wickham, H. (2015). *ggplot2: Elegant Graphics for Data Analysis*: Springer.

Witten, I. H.; Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques* (3rd ed.): Morgan Kaufmann Publishers.

Xu, R., & Wunsch, D. (2009). *Clustering*. New Jersey: Wiley-IEEE Press.