Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **3**, Number **1**, January **2018**

doi:10.3934/bdia.2018002

pp. **43–54**

AN APPLICATION OF PART TO THE FOOTBALL MANAGER DATA FOR PLAYERS CLUSTERS ANALYSES TO INFORM CLUB TEAM FORMATION

Marco Tosato and Jianhong Wu

Laboratory for Industrial and Applied Mathematics
York University
Toronto, Ontario, Canada, M3J 1P3

(Communicated by Xiaomin Zhu)

Abstract. We aim to show how a neural network based machine learning projective clustering algorithm, Projective Adaptive Resonance Theory (PART), can be effectively used to provide data-informed sports decisions. We illustrate this data-driven decision recommendation for AS Roma player market in the Summer 2018 season, using the two separate databases of fourty-seven attributes taken from Football Manager 2018 for each of the twenty-four soccer player, with the first including players of the AS Roma squad 2017-18, and the second consisting of all players linked with transfer moves to AS Roma. This is high dimensional data as players should be grouped only in terms of their performance with respect to a small subset of attributes. Projective clustering analyses provide a purely data-driven analysis to identify critical attributes and attribute characteristics for a group of players to form a natural cluster (in lower dimensional data space) in an unsupervised way. By merging the two databases, our unsupervised clustering analysis provides evidence-based recommendations about the club team formation, and in particular, the decision to buy and sell players within the same clusters, under different scenarios including financial constraints.

1. **Introduction.** High dimensional data arises naturally from the track performance records of players and some near-reality sports games. For example, the Football Manager, a managerial simulation platform and data set, has grown to be an integral part of soccer culture since 2014. Football Manager has been intensively analyzed and widely used by a large number of scouting network, as it has a worldwide database of the players and contains detailed history, statistics and attributes of all the included players. The recent deal signed with Prozone inglobates Football Manager's database to an authentic football analytics' firm and facilitates its diffusion to clubs globally [10].

Unsupervised clustering analyses of these sports data are both challenging and rewarding. Players have different roles in each team (and each game) which require different skill sets. Therefore, they normally form clusters with respect to different attributes in the data set, so projective clustering (finding clusters in lower dimensional subspaces of the data space) becomes necessary. How to mine these data

2010 *Mathematics Subject Classification.* Primary: 62-07; Secondary: 62P30, 62H30, 97M99.

Key words and phrases. PART, subspace clustering, unsupervised learning, sports analytics, football manager, soccer.

to identify respective attributes with respect to which clusters are formed and the cluster features emerge is an important research topics in sports analytics. Here, we use the analysis of Football Manager 2018 to show how a neural network based machine learning projective clustering algorithm PART provides an effective tool in mining this data to discover natural clusters of players. This analysis seems to be quite rewarding, as this provides a data and performance record-based analysis to detect players with certain similarities and hopefully, can help the manager to select players according to their needs. In particular, for the data set we are examining we hope our analysis can help AS Roma team manager Monchi to choose and decide how to act in the players' market at the beginning of the new season. It is hoped this data and record-based analysis supplements the professional AS Roma staff team in its decision making about which players should leave the team and which should be bought. This is important in particular for this season: even with some financial improvement in 2017 since AS Roma reached the semifinals of Champions League, there has been a deficit which needs to be settled and some money to be earned before the building of a new stadium [1]. So the consideration of balancing the budget through selling some of the most expensive players of the squad and replacing them with more affordable ones from the same player cluster becomes an option to maintain both the competitive level and the financial health (www.transfermarkt.com).

The rest of this paper is organized as follows: We first describe the data (Section 2), and then describe the PART algorithm with some discussions on how algorithm parameters should be chosen for the particular data we are examining (Section 3). We then report the clustering results and analyze some of the clustering features (Section 4), and make some recommendation on how team composition can be informed from this clustering analysis (Section 5).

2. Dataset.

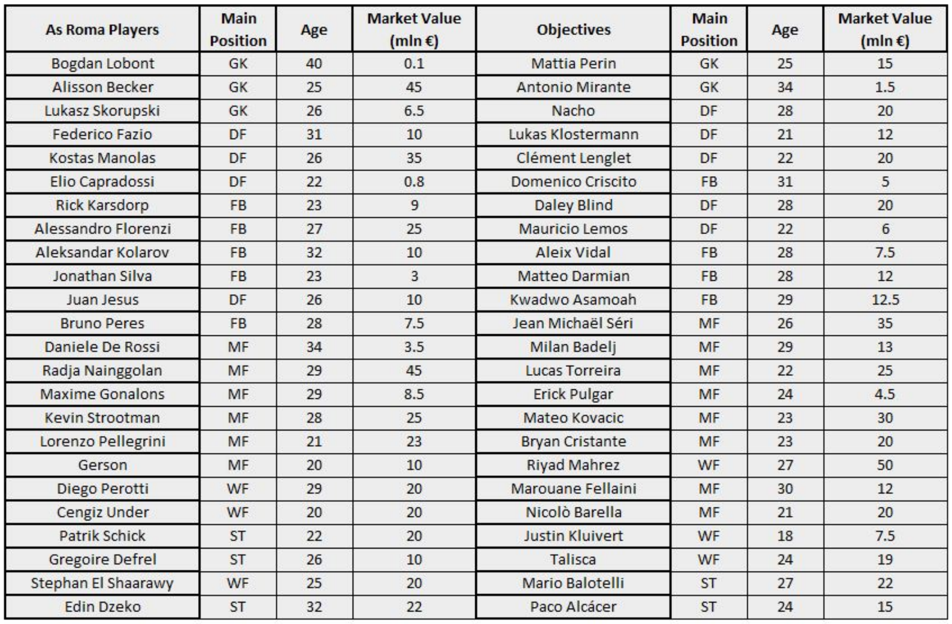
As mentioned above, the data we would like to analyze comes from a managerial simulation platform called Football Manager 2018. This data set has been intensively analyzed and widely used by soccer clubs, as it has a worldwide database of the players and contains detailed history, statistics and attributes of all the included players.

In the platform, each player (that will be considered as a vector in the PART clustering algorithm) is associated with certain attributes and characteristics, rated in ascending order from one to twenty, which are organized in different categories [2] as follows:

1. **Technical** - includes mostly abilities of players with the ball including set plays, passing, crossing, dribbling, finishing, heading, ball control (first touch), technique. Important defensive attributes to remark though are marking and tackling.
2. **Mental** - includes amongst all qualities related to:
    - **Commitment** - aggression, bravery, determination and work rate;
    - **Game knowledge and prediction** - anticipation, decision-making, flair, off-the-ball qualities, teamwork and vision;
    - **Personality** - composure and concentration.
3. **Physical** - includes amongst all qualities related to:
    - **Mobility** - acceleration, pace and stamina
    - **Structure** - body balance and strength.

4. **Goalkeeping -** includes saving skills which space from aerial ability to reflexes to one on ones and also teamwork and throw-in abilities.

Two databases will be analyzed: this includes the AS Roma first team squad of 2017-2018 and a list of twenty-four players which are considered as being potentially transferred to AS Roma (www.calciomercato.it; www.romanews.eu; www.tuttomercatoweb.com). These players are listed in Figure 1.



As Roma Players	Main Position	Age	Market Value (mln €)	Objectives	Main Position	Age	Market Value (mln €)
Bogdan Lobont	GK	40	0.1	Mattia Perin	GK	25	15
Alisson Becker	GK	25	45	Antonio Mirante	GK	34	1.5
Lukasz Skorupski	GK	26	6.5	Nacho	DF	28	20
Federico Fazio	DF	31	10	Lukas Klostermann	DF	21	12
Kostas Manolas	DF	26	35	Clément Lenglet	DF	22	20
Elio Capradossi	DF	22	0.8	Domenico Criscito	FB	31	5
Rick Karsdorp	FB	23	9	Daley Blind	DF	28	20
Alessandro Florenzi	FB	27	25	Mauricio Lemos	DF	22	6
Aleksandar Kolarov	FB	32	10	Aleix Vidal	FB	28	7.5
Jonathan Silva	FB	23	3	Matteo Darmian	FB	28	12
Juan Jesus	DF	26	10	Kwadwo Asamoah	FB	29	12.5
Bruno Peres	FB	28	7.5	Jean Michaël Séri	MF	26	35
Daniele De Rossi	MF	34	3.5	Milan Badelj	MF	29	13
Radja Nainggolan	MF	29	45	Lucas Torreira	MF	22	25
Maxime Gonalons	MF	29	8.5	Erick Pulgar	MF	24	4.5
Kevin Strootman	MF	28	25	Mateo Kovacic	MF	23	30
Lorenzo Pellegrini	MF	21	23	Bryan Cristante	MF	23	20
Gerson	MF	20	10	Riyad Mahrez	WF	27	50
Diego Perotti	WF	29	20	Marouane Fellaini	MF	30	12
Cengiz Under	WF	20	20	Nicolò Barella	MF	21	20
Patrik Schick	ST	22	20	Justin Kluivert	WF	18	7.5
Gregoire Defrel	ST	26	10	Talisca	WF	24	19
Stephan El Shaarawy	WF	25	20	Mario Balotelli	ST	27	22
Edin Dzeko	ST	32	22	Paco Alcácer	ST	24	15

FIGURE 1. List of players from the two data sets to be examined: the list contains information of their ages and market values.

3. **The PART algorithm.** We will use the Projective Adaptive Resonance Theory (PART) algorithm for clustering analysis. PART is a neural network based clustering algorithm developed in [5, 6] that has been shown to be efficient to sort out high-dimensional data and identify patterns embedded in lower dimensional subspaces and has already been applied in different areas including: social media [14], neural spiking trains [9] and gene filtering for cancer diagnosis [13, 12].

The use of PART to our datasets is natural: the dimension of the data is the number of characteristics considered for each player. These datasets in fourty-seven dimensional space provide challenges for traditional clustering procedures such as k-means as subspaces where clusters are formed are not given in advance. Finding the clusters and the respective dimensions where clusters are formed should be processed in parallel, and PART conducts subspace clustering in such an unsupervised way that it selects and rejects relevant dimensions through adaptive learning to discover and underline important characteristics of players in an emerging player cluster. For example, PART will identify the skill of crossing and dribbling as key characteristics in a cluster most likely associated with wingers, defensive skills instead for a center back cluster.

3.1. **PART components and parameters.** In the PART algorithm, the index i referees to the dimension of the data in the input layer, j referees to cluster in the clustering layer, m is the total number of dimensions, which in our study, is fourty-seven; and n is the total number of vectors to cluster and this is the maximal number of clusters, obviously.

The algorithm developed in the series of papers [8, 5, 6], is made up of four main steps and requires several parameters discussed below:

1. *F_1 activation and computation of selective output signals h_{ij}*: here there are two parameters that need to be initialized by the users: ζ is a threshold that determines if a dimension is relevant or not for the cluster even there is a similarity between the input data and the potential cluster with respect to that particular dimension; and σ is a distance vigilance parameter which determines how far apart can two attribute values should be for them to belong to two different clusters which are formed with respect to the same attribute. We will discuss a procedure on how σ should be initialized and refined for our datasets. The algorithm requires the calculation of the similarity measure h_{ij} defined as follows:

   $$h_{ij} = \begin{cases} 1 & \text{if } |x_i - z_{ji}| \leq \sigma \wedge z_{ij} > \zeta \\ 0 & \text{otherwise.} \end{cases}$$

   This gives a matrix of dimensions $n \times m$, where h_{ij} is a binary value (either zero or one) which checks whether the value in dimension i of the vector is close enough to the feature of the representative of the j-th cluster. This is in relevant to z_{ij}, the bottom-up weight, that measures the significance of dimension i if the input is assigned to the j-th cluster.
2. *F_2 activation and selection of winner*: To select the winner cluster for a given input, PART proceeds to compute the bottom-up filter input T_j to committed F_2 node v_j using the formula below:

   $$T_j = \sum_i z_{ij} h_{ij}$$

   This takes into consideration the fraction of significant dimensions which display similarity between the new vector and the cluster j. If there are no committed v_j, then the input vector is added and forms a new cluster, otherwise choose the maximum of T_j's in a winner-takes-all paradigm.
3. *Vigilance and reset*: The vigilance parameter, ρ, is another user-specified parameter. This is the minimum number of dimensions with the needed similarity between a vector input and the cluster characteristics, in order for the vector to be part of the cluster. When

   $$\sum_i h_{ij} < \rho,$$

   the winning node is reset as the number of attributes to allocate this input vector to a cluster is not sufficiently large.

4. *Learning*: If the j-th cluster node is a committed winning node, then, given the set $X = \{i : h_{ij} = 1\}$ of relevant dimensions, we have the learning rules:
$$z_{ij}^{\text{new}} = \begin{cases} L/(L-1+|X|) & \text{if } h_{ij} = 1 \\ 0 & \text{otherwise.} \end{cases}$$
and
$$z_{ji}^{\text{new}} = (1-\alpha)\, z_{ji}^{\text{old}} + \alpha x_i.$$
If, on the other hand, is the j-th node is not a committed node, then we have weights for the new cluster calculated as:
$$z_{ij}^{\text{new}} = \frac{L}{L-1+m}$$
and
$$z_{ji}^{\text{new}} = x_i,$$
where α is a learning rate between 0 and 1 which quantifies the importance to be given to the new vector inside the cluster. This is the other parameter to be specified by the user. In the second case, a new categorization is formed with the initial set of characteristics determined by the above rule from the first learning experience.

In practice, the above procedure is repeated for the entire list of vectors (in this case for all players included in the two data sets) multiple times until the top-down weights of each cluster are stable (if the maximum distance is smaller than a small tolerance) between subsequent iterations. An illustration of the algorithm just explained is given in Figure 2.

3.2. **Initiation and adaptation of algorithm parameters and measure of clustering quality.** The two databases are fed in the PART algorithm, where every player represents an input vector and his characteristics are the values of the corresponding attributes. The idea is to perform an unsupervised clustering to check similarities between players which reflect their positions in the field (for example, in the outcome of clustering analysis, it is expected for goalkeepers to be all in the same cluster, and not misplaced in other clusters).

To achieve this, we need to tune the choice of the parameters in the PART algorithm. For an optimal process, we need to introduce two similarity measures:

- Within Cluster Similarity Measure (WCSM): which calculates the global average distance in relevant characteristics between vectors that are part of the same cluster; the smaller the WCSM, the better the clustering results;
- Between Cluster Distance (BCD): which calculates the average distance in relevant characteristics between the different clusters; the larger the BCD, the better the clustering results.

Let $\tilde{N}$ be the number of clusters with more than one vector (total number is N), S_j be set of relevant dimensions for cluster j which has n_j vectors with $\bar{X}_i$ being the average of points in the cluster. We define
$$WCSM = \frac{1}{\tilde{N}} \sum_{j \in N} \sum_{i \in S_j} \frac{|X_i - \bar{X}_i|}{n_j};$$
and
$$BCD = \frac{1}{N^2 - N} \sum_{j_1 \in N} \sum_{j_2 \in N} \sum_{i \in S_{j_1} \cap S_{j_2}} |X_{i_{j_2}} - X_{i_{j_1}}|.$$

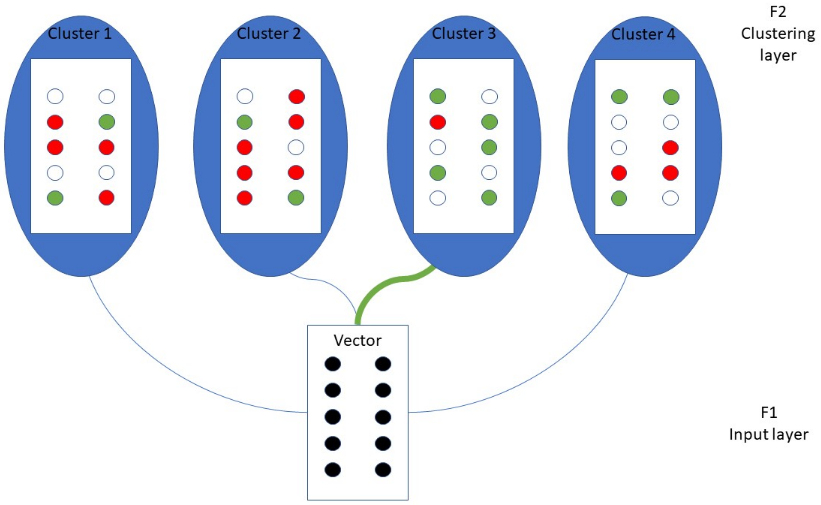


FIGURE 2. An illustration of vector clustering using PART where each ball represents a dimension of the vector. At first there is a similarity check where the new vector is compared to the representative of the four clusters. Empty balls show irrelevant dimensions for the cluster ($z_{ij} \leq \zeta$) differently from full-coloured balls ($z_{ij} > \zeta$). This group can be subdivided further into dimensions in which new vector is similar enough to the respective top-down weight z_{ji} (distance is smaller than σ) depicted in green and when they are well separated (distance is greater than σ) depicted in red.
In this case, the algorithm starts by comparing the cluster that yields the highest similarity by computing the T_j's as shown above, in this case cluster 3, and counts the number of similar relevant dimensions (green balls) to compare it with ρ. If $\rho \leq 5$ the vector is included in the third cluster; otherwise (if $\rho > 5$) it is compared to all other clusters ordered based on their T_j's; in this case since in every cluster the green balls are no more than five, the new vector cannot be categorized in the existing clusters and will form a new one of its own.
Moreover, in the first case ($\rho \leq 5$), the representative of cluster 3 will be updated by taking into consideration $\alpha\%$ of the new value and the relevant dimensions will become only those relative to the green balls; in the second case it will form a new cluster where every dimension has equal importance and the representative of the fifth cluster formed would be the vector itself.

So, the goal is to minimize

$$CM = \frac{WCSM}{BCD}.$$

Note that in the case where each vector belongs to one and only one cluster, we have $CM = 0$. Therefore, we need to decide in advance what would be the acceptable number of clusters so that the optimization can stop. Here we choose the maximal number of clusters as eight since this seems to present the various positions in soccer (from goalkeeper to striker).

Given a maximum number of clusters, the parameters will then be chosen according to a grid where the cluster measure CM can be minimized. In particular, we have

$$\sigma \in \{1, 2, 3\}; \quad \rho \in \{16, 17, 18, 19, 20\}; \quad \alpha \in \{0.1, 0.2, 0.3\}.$$

4. Cluster results and club team composition recommendations.

The clustering results solely based on the application of PART algorithm to the two datasets have clearly identified interesting characteristics of different clusters. We now describe the clustering results for the first data set and the second data separately, and then to the merged single data set.

4.1. Clustering results for data sets separately.

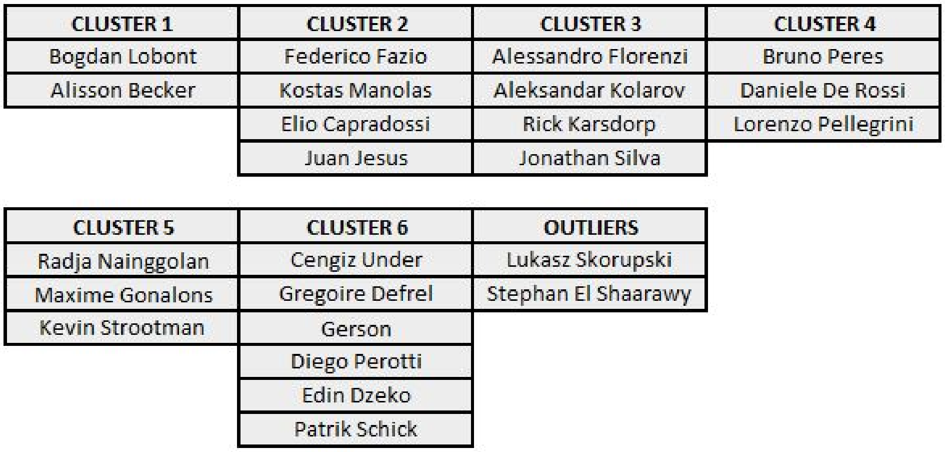
1. *Dataset 1*: In this first data set, we observe that players are quite precisely sorted according to their positions on the field and the relevant characteristics per cluster are closely related to the position in the field.

   In particular, cluster 1 includes just two goalkeepers out of three and it is interesting to note that the goalkeeper Skorupski is not included here since he is younger and less experienced than the other two and so is mostly lacking in some mental attributes which include concentration, decision, determination and also communication. Important to note is that in cases in which the distance $\sigma = 3$, Skorupski is correctly identified with the other goalkeepers.

   Cluster 2 includes all center-backs which have high marking, positioning and teamwork qualities; cluster 3 includes full-backs which have high dribbling, passing, agility and most commitment and personality mental qualities; cluster 5 includes box-to-box central midfielders which mostly share a good ball handling and game knowledge and prediction mental attributes such as anticipation, decision, positioning and vision.

   Cluster 6 is predictably made up of most of AS Roma's offensive players which have in common certain commitment mental attributes such as aggression and bravery and also strength. Note that we would expect El Shaarawy also to be in this cluster but this does not occur since he does not match some of the mental attributes in the cluster such as bravery and he has overall less strength than the rest of the players in this cluster.

   The remaining cluster - cluster 4 is unexpected since it is made up of polyvalent players which, at first sight, do not seem to have much in common and contains: the full-back or winger Bruno Peres, the AS Roma captain De Rossi and the polyvalent youngster Lorenzo Pellegrini. The common characteristics for this cluster are mostly technical including crossing, first touch, marking and technique and physical including body balance and stamina but present completely different mental attributes. This would be expected since De Rossi is the leader and captain of the team and has a completely different personality to the funambolic brazilian Bruno Peres.



CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
Bogdan Lobont	Federico Fazio	Alessandro Florenzi	Bruno Peres
Alisson Becker	Kostas Manolas	Aleksandar Kolarov	Daniele De Rossi
	Elio Capradossi	Rick Karsdorp	Lorenzo Pellegrini
	Juan Jesus	Jonathan Silva	

CLUSTER 5	CLUSTER 6	OUTLIERS
Radja Nainggolan	Cengiz Under	Lukasz Skorupski
Maxime Gonalons	Gregoire Defrel	Stephan El Shaarawy
Kevin Strootman	Gerson	
	Diego Perotti	
	Edin Dzeko	
	Patrik Schick	

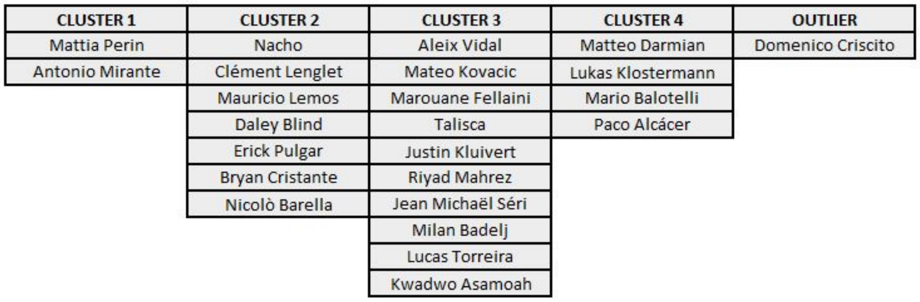
Figure 3. Unsupervised clustering of AS Roma first team players 2017/2018 ($\sigma = 2;\ \rho = 18;\ \alpha = 0.2$)

2. *Dataset 2*: In the second dataset, optimizing the cluster measure results in a smaller amount of clusters (four) since the parameters present a rather high value of σ (three) which means that characteristics between different players are considered similar in all cases in which they differ for three or less while in previous case it was two; moreover $\rho = 17$ is also smaller than previous case, thus just seventeen similar characteristics are needed to be in the same cluster.

   In this case, Cluster 1 represents both the goalkeepers of the data set which have most of the characteristics in common (thirty-five out of forty-seven) and the main different goalkeeping skills include communication and also coming out of the goal; cluster 2 contains players in defensive positions including centre-backs and holding midfielders which have certain defensive technical attributes in common such as heading and tackling and have a high body balance.

   Cluster 3 is made up of a variety of more offensive players and include a lot of central midfielders which have as main common features ball control which includes first touch and technique and also some physical mobility attributes. The most unexpected cluster is the fourth which is made up of two defenders and two main strikers; since those players had different goals in the field we see that the similarities occur in the mental and physical range including aggression, anticipation, concentration, body balance and stamina while there is no common technical attribute.

   Important to note is that Domenico Criscito in this simulation is an outlier and this is not surprising since this player presents some uncommon characteristics for a full back, for example he is good in penalty taking. Also in the following simulation, we will also see that Criscito is not placed with other players in his position and instead is included with many midfielders.



CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	OUTLIER
Mattia Perin	Nacho	Aleix Vidal	Matteo Darmian	Domenico Criscito
Antonio Mirante	Clément Lenglet	Mateo Kovacic	Lukas Klostermann	
	Mauricio Lemos	Marouane Fellaini	Mario Balotelli	
	Daley Blind	Talisca	Paco Alcácer	
	Erick Pulgar	Justin Kluivert		
	Bryan Cristante	Riyad Mahrez		
	Nicolò Barella	Jean Michaël Séri		
		Milan Badelj		
		Lucas Torreira		
		Kwadwo Asamoah		

FIGURE 4. Unsupervised clustering of AS Roma market objectives for Summer 2018 ($\sigma = 3;\ \rho = 17;\ \alpha = 0.2$)

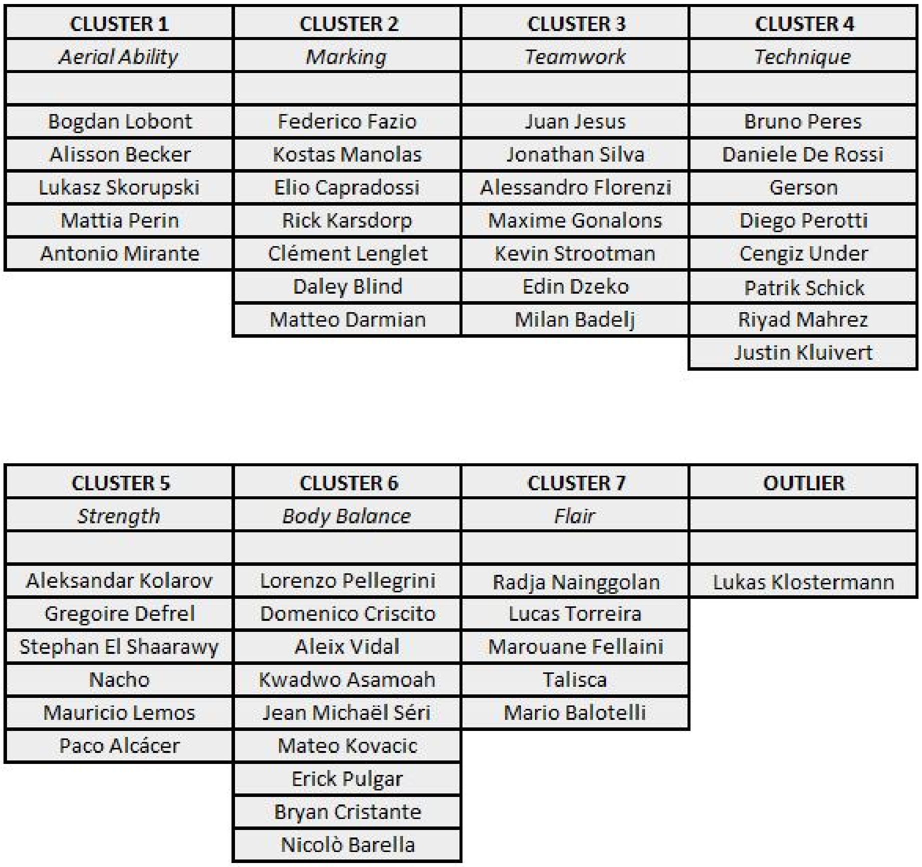
3. *A Single Merged Dataset*: As in the previous clusterizations, we look for the most important attributes correlated to each cluster but this time will use them to try and help in making a decision about which players to buy and sell, and these players should be from the same cluster.

Starting from the most expected clusters: Cluster 1 includes all goalkeepers as expected which have similar aerial ability, command of the area and throw-in abilities, Cluster 2 groups all players which play in defense and have expected characteristics to reflect their main strength such as marking, anticipation and concentration. Interesting to note is the fact that in this group of players includes one of AS Roma's most valuable player Manolas and one of the most promising youngsters of French football: Clement Lenglet.

There is also a more technical group of players which is classified in Cluster 4 and mostly contains wingers in addition to the polyvalent De Rossi where their main common attributes are technique, passing and ball control as expected but also mobility-related attributes such as pace and stamina; and a more physical group of players contained in Cluster 5 made up of strong, powerful defenders and also strikers where the common attributes are mostly physical structure dependent - body balance and strength and related to commitment - bravery and determination.

Most of the central midfielders are in Cluster 6, to which are added two other players: Aleix Vidal (which has also played winger in his career) and especially Domenico Criscito which presents characteristics uncommon for full-backs and are difficult to categorize. All these players share similarities in almost half the characteristics (twenty-three) and the relevant ones include tackling, most mental attributes including leadership, teamwork and commitment-related and also acceleration and pace as physical attributes.

Finally the most unpredictable clusters are Cluster 3 which contains a lot of thoughtful team players ranging from central defenders like Juan Jesus to strikers like Edin Dzeko who have minimum number of common characteristics (twenty) which include dribbling, passing, aggression and positioning; and Cluster 7 which presents really impactful players in the game that are able to change it with their leading qualities and include players like Radja Nainggolan and Lucas Torreira. This last cluster is really interesting since the players here present similar characteristics throughout the entire range of technical



CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
Aerial Ability	*Marking*	*Teamwork*	*Technique*
Bogdan Lobont	Federico Fazio	Juan Jesus	Bruno Peres
Alisson Becker	Kostas Manolas	Jonathan Silva	Daniele De Rossi
Lukasz Skorupski	Elio Capradossi	Alessandro Florenzi	Gerson
Mattia Perin	Rick Karsdorp	Maxime Gonalons	Diego Perotti
Antonio Mirante	Clément Lenglet	Kevin Strootman	Cengiz Under
	Daley Blind	Edin Dzeko	Patrik Schick
	Matteo Darmian	Milan Badelj	Riyad Mahrez
			Justin Kluivert

CLUSTER 5	CLUSTER 6	CLUSTER 7	OUTLIER
Strength	*Body Balance*	*Flair*	
Aleksandar Kolarov	Lorenzo Pellegrini	Radja Nainggolan	Lukas Klostermann
Gregoire Defrel	Domenico Criscito	Lucas Torreira	
Stephan El Shaarawy	Aleix Vidal	Marouane Fellaini	
Nacho	Kwadwo Asamoah	Talisca	
Mauricio Lemos	Jean Michaël Séri	Mario Balotelli	
Paco Alcácer	Mateo Kovacic		
	Erick Pulgar		
	Bryan Cristante		
	Nicolò Barella		

Figure 5. Unsupervised clustering of the two data sets merged as a single data, with underlined indicative characteristic per cluster ($\sigma = 3$; $\rho = 20$; $\alpha = 0.2$)

(passing, technique and first touch play), mental (anticipation, composure, decisions and flair) and physical skills (agility and pace).

In short, we conclude that an application PART to the two data sets provide insights about what underlines the main and common characteristics of players to make them be part of a particular cluster. The application of PART also identifies meaningful outliner: Lukas Klostermann remained uncategorized since he presented different abilities than all the other player groups identified by this neural network clustering algorithm.

4.2. **Team formation.** Taking into consideration the age and market values of the players considered and in terms of some critical characteristics shared by the same cluster of players identified by the PART algorithm, we hypothesize two possible exchanges in the 2018 Summer market so that AS Roma would be able to resettle its finances and not lose certain characteristics:

- Sell Radja Nainggolan who is thirty years old and has a market value of fourty-five million euros; and replace him with Lucas Torreira who is much younger (twenty-two years old), has a bright future, costs about twenty million euros less than Nainggolan; and most importantly shares (according to PART) a wide range of important characteristics including on-play (technical) abilities, game knowledge and personality (mental) and also mobility (physical) attributes.
- Sell Konstantinos Manolas who is 26 years old and has a market value of 35 million euros, and replace him with Clement Lenglet who is a youngster (22 years old) and has been one of the revelations in La Liga last season playing for Seville, and costs about 15 million euros less than the former, while maintaining similar defensive skills as shown through the clustering analysis using PART.

5. **Remarks.** We have introduced the projective clustering algorithm PART, and illustrated its effectiveness in generating meaningful projective clusters of players from two important data sets from the Football Manager platform. The PART algorithm, parametrized from knowledge about the data and the desire of clustering results for team formation recommendation, categorized relevant players into clusters with common characteristics that defines a cluster and distinct the cluster from others. A direct application of the clustering results was the data-driven recommendation on which players can be substituted. We conclude that PART as a neural network based machine learning clustering algorithm has the potential to be an effective decision support Sports Analytics tool.

Other examples of subspace algorithms can also potentially be applied to this dataset [11]. These algorithms include Bottom-up Subspace Search methods which use an Apriori style approach and include CLIQUE (CLustering In QUEst)[4] and ENCLUS (ENtropy-based CLUStering) [7]; and Iterative Top-down Subspace Search methods where weights are assigned and updated to each dimension in the cluster and include PROCLUS (PROjective CLUStering) [3]. It is an interesting project for a future study to compare the clustering results using all available high-dimensional data clustering algorithms.

Acknowledgments. This work has been partially supported by NSERC Discovery Grant.

REFERENCES

[1] Financial Fair Play club summary - Official document, retrieved from www.uefa.com.

[2] Guide to Football Manager, retrieved from www.guidetofm.com.

[3] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc and J. S. Park, Fast algorithms for projected clustering, *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, ACM Press, (1999), 61–72.

[4] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications, *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, ACM Press (1998), 94–105.

[5] J. Cao and J. Wu, Dynamics of projective adaptive resonance theory model: The foundation of PART algorithm, *IEEE Transactions on Neural Networks*, 2004.

[6] J. Cao and J. Wu, Projective ART for clustering data sets in high dimensional spaces, *Neural Networks*, 2002.

[7] C. H. Cheng, A. W. Fu and Y. Zhang, Entropy-based subspace clustering for mining numerical data, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press (1999), 84–93.

[8] G. Gan, C. Ma and J. Wu, Data clustering: Theory, algorithms and applications, *SIAM, Philadelphia, ASA, Alexandria, VA,* (2007), xxii+466 pp.

[9] J. D. Hunter, J. Wu and J. Milton, Clustering neural spike trains with transient responses, *Proceedings of the 47th IEEE Conference on Decision and Control*, 2008.

[10] Director - L. Myles, An alternative reality: The football manager documentary, *Sports Interactive*, 2014.

[11] L. Parsons, E. Haque and H. Liu, Subspace clustering for high dimensional data: A review, *ACM*, 2004.

[12] H. Takahashi and H. Honda, Modified signal-to-noise: a new simple and practical gene filtering approach based on the concept of projective adaptive resonance theory (PART) filtering method, *Bioinformatics*, 2006.

[13] H. Takahashi, T. Kobayashi and H. Honda, Construction of robust prognostic predictors by using projective adaptive resonance theory as a gene filtering method, *Bioinformatics*, 2005.

[14] J. Wu, Projective adaptive resonance theory revisited with applications to clustering influence spread in online social networks, *Data Analytics*, 2015.

E-mail address: halohalo@mathstat.yorku.ca
E-mail address: wujh@mathstat.yorku.ca