Clustering Professional Baseball Players with SOM and Deciding Team Reinforcement Strategy with AHP

Kazuhiro Kohara(✉) and Shota Enomoto

Chiba Institute of Technology, 2-17-1 Tsudanuma,
Narashino, Chiba 275-0016, Japan
kohara.kazuhiro@it-chiba.ac.jp

Abstract. In this paper, we propose an integration method that uses self-organizing maps (SOM) and the analytic hierarchy process (AHP) to cluster professional baseball players and to make decision on team reinforcement strategy. We used data of pitchers in the Japanese professional baseball teams. First, we collected data of 302 pitchers and clustered these pitchers using the following fourteen features: number of games pitched, number of wins, number of loses, number of save, number of hold, number of innings pitched, rate of strikeout, ERA (earned run average), percentage of hits a pitcher allows, WHIP (walks plus hits per inning pitched), K/BB (strikeout to walk ratio), FIP (fielding independent pitching), LOB% (left on base percentage), RSAA (runs saved above average). Second, we created pitcher maps of all teams and each team with SOM. Third, we examined main features of each cluster. Fourth, we considered team reinforcement strategies by using the pitcher maps. Finally, we used AHP to determine the team reinforcement strategy.

Keywords: Clustering · Visualization · Data mining
Business intelligence · Sport industry · Baseball · Decision making
Self-organizing maps · AHP

1 Introduction

Machine learning and data mining techniques have been extensively investigated, and various attempts have been made to apply them to baseball e.g., [1–5]. Tolbert and Trafalis applied SVM (Support Vector Machine) to predicting MLB (Major League Baseball) championship winners [1]. Ishii applied K-means clustering to identifying undervalued baseball players [2]. Pane applied K-means clustering and Fisher-wise criterion to identifying clusters of MLB pitchers [3]. Tung applied PCA (Principal Component Analysis) and K-means clustering to analyzing a multivariate data set of career batting performances in MLB [4]. Vazquez applied time series and clustering algorithms to predicting baseball results [5]. In this paper, we propose an integration method that uses Self-Organizing Maps (SOM) [6] and the analytic hierarchy process (AHP) [7] to cluster professional baseball players and to make decision on team reinforcement strategy. We used data of pitchers in Japanese baseball teams. First, we collected data of 302 pitchers and clustered these pitchers using fourteen features. Second,

P. Perner (Ed.): ICDM 2018, LNAI 10933, pp. 135–147, 2018.
https://doi.org/10.1007/978-3-319-95786-9_10

we created pitcher maps of all teams and each team with SOM. Third, we examined main features of each cluster. Fourth, we considered team reinforcement strategies by using pitcher maps. Finally, we used AHP to determine the team reinforcement strategy.

2 Clustering Professional Baseball Players with SOM

The SOM algorithm is based on unsupervised, competitive learning [6]. It provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice and thus the mapping is a mapping from high dimensional space onto a plane.

Previously, we proposed a way of purchase decision support using SOM and AHP. First, we provided two class boundaries, which divide the range between the maximum and minimum of an input feature value into three equal parts. Second, we created self-organizing product maps using the classified inputs. We applied our way to five kinds of products and confirmed its effectiveness [8]. When we previously compared SOM with the other clustering algorithms (hierarchical clustering and K-means clustering) for product clustering, SOM were superior to the other clustering algorithms for both visibility and clustering ability [9]. Therefore, we used SOM for baseball players clustering.

We used data of pitchers of NPB (Nippon Professional Baseball Organization) [10]. We collected data of 302 pitchers in 2015 from Japanese professional baseball database [10, 11]. We clustered these pitchers using the following fourteen features: number of games pitched, number of wins, number of loses, number of save, number of hold, number of innings pitched, rate of strikeouts, ERA (earned run average), percentage of hits a pitcher allows, WHIP (walks plus hits per inning pitched), K/BB (strikeout to walk ratio), FIP (fielding independent pitching), LOB% (left on base percentage), RSAA (runs saved above average).

In each feature, we provide two class boundaries, which divide the range between the maximum and minimum of an input feature value into three equal parts. For classifying the data of the number of games pitched, we divided the number into three classes: under 27, over 28 to 50, and over 51. For classifying the data of the number of wins, we divided the number into three classes: under 5, over 6 to 10, and over 11. For classifying the data of the number of loses, we divided the number into three classes: under 4, over 5 to 8, and over 9. For classifying the data of the number of save, we divided the number into three classes: under 13, over 14 to 27, and over 28. For classifying the data of the number of hold, we divided the number into three classes: under 13, over 14 to 26, and over 27. For classifying the data of the number of innings pitched, we divided the number into three classes: under 74, over 75 to 140, and over 141. For classifying the data of the rate of strikeouts, we divided the rate into three classes: under 6.09, over 6.10 to 10.15, and over 10.16. For classifying the data of ERA, we divided ERA into three classes: under 3.52, over 3.53 to 6.64, and over 6.65. For classifying the data of the percentage of hits a pitcher allows, we divided the percentage into three classes: under 8.35, over 8.36 to 13.08, and over 13.09. For classifying the data of WHIP, we divided WHIP into three classes: under 1.36, over 1.37 to 2.08, and over 2.09. For classifying the data of K/BB, we divided K/BB into three classes: under 4.70, over 4.71 to 8.85, and over 8.86. For classifying the data of FIP, we divided FIP into three classes: under 3.20, over 3.21 to 5.27, and over 5.28.

For classifying the data of LOB%, we divided LOB% into three classes: under 0.661, over 0.662 to 0.814, and over 0.815. For classifying the data of RSAA, we divided RSAA into three classes: under −2.1083, over −2.1082 to 16.65, and over 16.66.

Table 1 shows a part of the feature matrix for pitchers.

Table 1. A part of the feature matrix for pitchers.

Name	Number of games pitched			Number of wins		
	Under 27	Over 28 to 50	Over 51	Under 5	Over 6 to 10	Over 11
Makita	0	1	0	0	1	0
Hamada	1	0	0	1	0	0
Sawamura	0	0	1	0	1	0
Settu	1	0	0	0	0	1
Wakui	0	1	0	0	0	1
Arihara	1	0	0	0	1	0
Masui	0	0	1	1	0	0
Fujinami	0	1	0	0	0	1
Yamaguchi	0	0	1	1	0	0

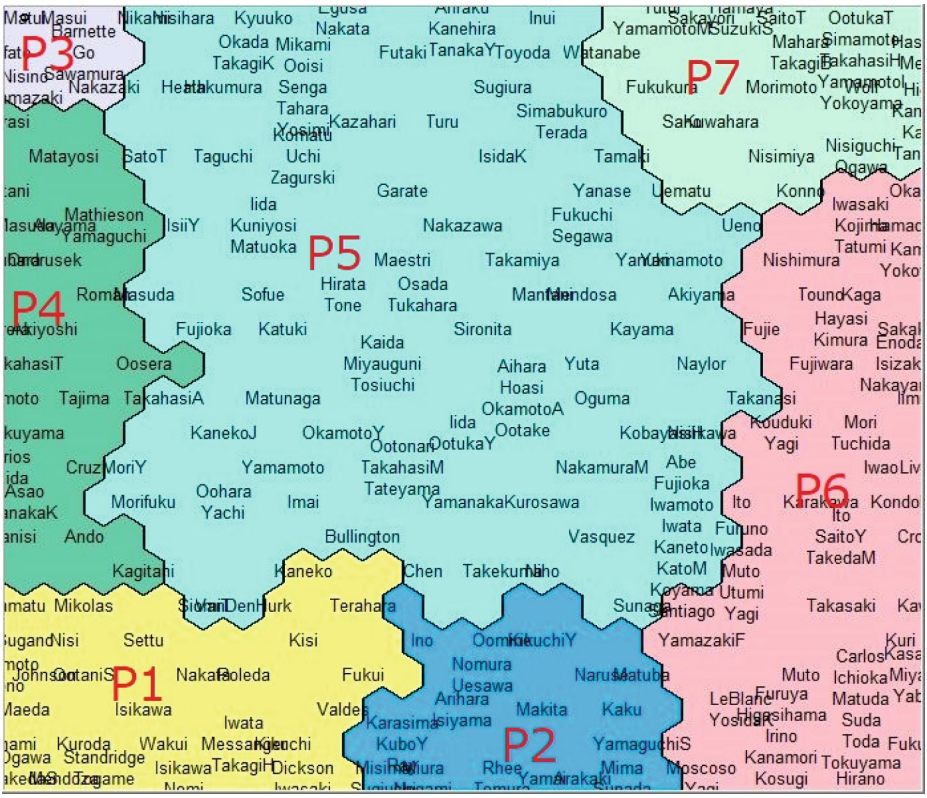
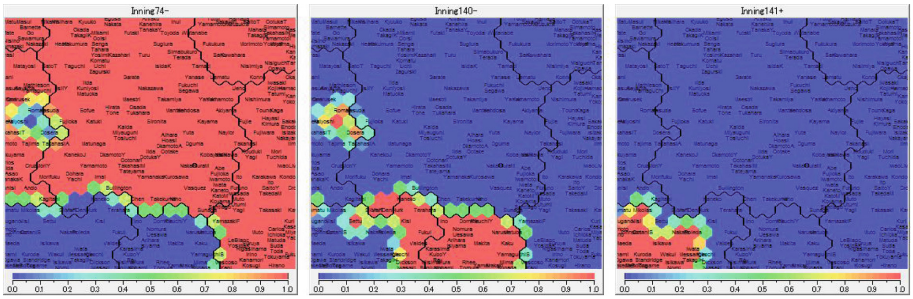


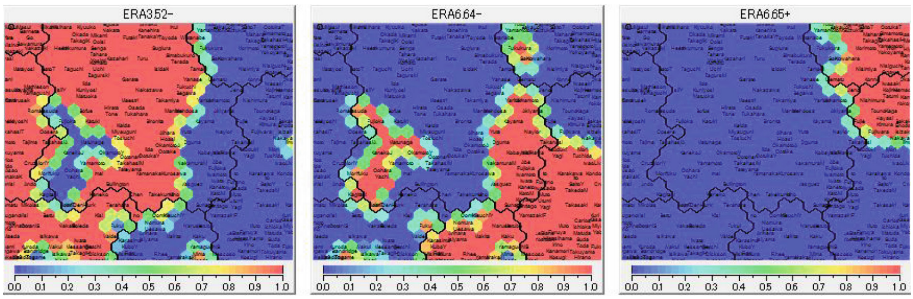
Fig. 1. Self-organizing cluster map of pitchers of all teams.

We inputted the data of all pitchers into SOM and created pitcher maps of all teams. Figure 1 shows self-organizing map of pitchers of all teams. Figures 2, 3 and 4 show examples of component maps of pitchers.



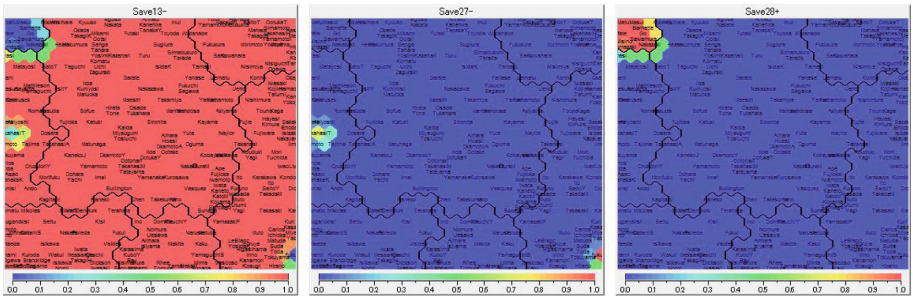
(a) under 74 (b) over 75 to 140 (c) over 141

Fig. 2. Component map of pitchers of all teams: number of innings pitched. (Color figure online)



(a) under 3.52 (b) over 3.53 to 6.64 (c) over 6.65

Fig. 3. Component map of pitchers of all teams: ERA (Earned Run Average). (Color figure online)



(a) under 13 (b) over 14 to 27 (c) over 28

Fig. 4. Component map of pitchers of all teams: number of save. (Color figure online)

There were seven clusters in Fig. 1. When inspecting component maps, the feature of each cluster is clear. For example, red neurons correspond to over 141 innings pitched in Fig. 2(c) and red neurons correspond to over 75 to 140 innings pitched in Fig. 2(b). Red neurons correspond to under 3.52 ERA in Fig. 3(a) and red neurons correspond to over 3.53 to 6.64 ERA in Fig. 3(b). Red neurons correspond to over 28 save in Fig. 4(c).

As the number of innings pitched is large (over 141) and ERA is small (under 3.52) in cluster P1, a pitcher belonging to P1 is one of the best starting pitcher. As the number of innings pitched is medium (over 75 to 140) and ERA is medium (over 3.53 to 6.64) in cluster P2, a pitcher belonging to P2 is one of the second best starting pitcher. As the number of save is large (over 28) and ERA is small (under 3.52) in cluster P3, a pitcher belonging to P3 is a *closer*. We inspected every component maps and understand that features of Clusters P1 to P7 are as shown in Table 2.

Table 2. Main features of all NPB pitchers in 2015 in each cluster.

Cluster	Features	Main feature
P1	Number of innings pitched is large Both ERA and WHIP are small	Best starting pitchers
P2	Number of innings pitched is medium Both ERA and WHIP are medium	Second best starting pitchers
P3	Number of save is large	Closer
P4	Number of hold is large	Best setup pitchers
P5	Number of wins and loses is small	Third best starting pitchers or second best setup pitchers
P6	Number of wins and loses is small Both ERA and WHIP are large	Fourth best starting pitchers or third best setup pitchers
P7	Number of innings pitched is small Both ERA and WHIP are large	Bad pitchers

3 Considering Team Reinforcement Strategies

Next, we inputted the data of pitchers belonging to Chiba Lotte Marines into SOM and created pitcher maps. Figure 5 shows self-organizing pitcher maps of Lotte.

We inspected every component maps and understand that main feature of Clusters L1 to L6 are as shown in Table 3.

Here, we assumed that organization of pitchers in a strong team is as follows: the number of starting pitchers is five to six, the number of setup pitchers is one to two, the number of closer is one to two, and the number of relief pitchers is three to five.

When comparing Lotte's organization of pitchers with a strong team's organization, we understand that the number of starting pitchers is not enough.

Here, we chose alternatives for reinforcement strategies of starting pitchers as follows.

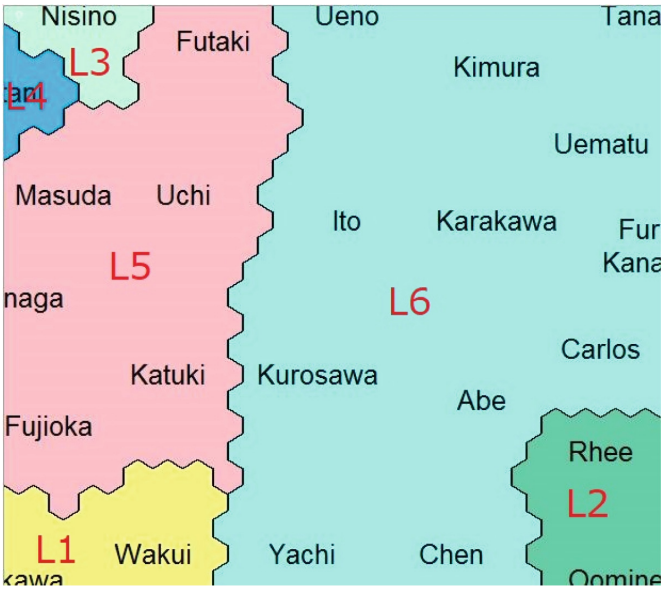


Fig. 5. Self-organizing cluster map of pitchers of Chiba Lotte Marines.

Table 3. Main features of pitchers of Chiba Lotte Marines in 2015 in each cluster.

Cluster	Main feature (# of pitchers)	Name (cluster in all NPB pitchers)
L1	Best starting pitchers (2)	Wakui, Isikawa (P1)
L2	Second best starting pitchers (2)	Rhee, Oomine (P2)
L3	Closer (1)	Nisino (P3)
L4	Best setup pitchers (1)	Ootani (P4)
L5	Second best setup pitchers (6)	Masuda, Fujioka, Matunaga, Uchi, Niki (P5), katuki (P6)
L6	Substitutes (13)	Kurosawa, Yachi, Abe, Chen (P5), Kanamori, Furuya, Carlos (P6)

Step 1: We choose pitchers (1) who belong to Clusters P1, P2, P5 or P6, (2) whose contract have been expired or who declared *free agent*, and (3) whose number of innings pitched was large or whose percentage of hits he allows was small. We chose Stanridge and Bullington.
Step 2: We choose pitchers (1) who belong to Clusters P1, P2, P5 or P6, (2) who are young and whose salary is low, (3) whose number of innings pitched was medium or whose FIP was small or whose RSAA was not small. We chose Iida and Mima.

Table 4 shows the data of four alternatives for a reinforced starting pitcher.

Table 4. Data of alternatives for a reinforced starting pitcher.

Name	Innings pitched	Hit ratio	RSAA	FIP	Salary (million yen)	Age	Right/ left
Standridge	144.3	9.4	–0.52	3.79	200	37	right
Bullington	73.6	7.3	2.378	3.18	150	35	right
Mima	86.3	10.6	–7.035	3.53	40	29	right
Iida	41.3	6.5	2.169	3.19	4	24	left

Hit ratio: percentage of hits a pitcher allows,
Right/left: right throw or left throw.

4 Decision Making on Team Reinforcement Strategy with AHP

AHP is a multi-criteria decision method that uses hierarchical structures to represent a problem [7]. Pairwise comparisons are based on forming a judgment between two particular elements rather than attempting to prioritize an entire list of elements. The AHP scales of pairwise comparisons are shown in Table 5.

Table 5. The AHP scales for pairwise comparisons.

Intensity of importance	Definition and explanation
1	Equal importance
3	Moderate importance
5	Essential or strong importance
7	Demonstrated importance
9	Extreme importance
2, 4, 6, 8	Intermediate values between the two adjacent judgments when compromise is needed

Figure 6 shows an example of the relative measurement AHP model created for the task of deciding a high capable pitcher. Here, we used the following four criteria: innings pitched, hit ratio (percentage of hits a pitcher allows), RSAA and FIP.

We assumed the pairwise comparison matrix for Ciba Lotte Marines. The pairwise comparison matrix for the four criteria is shown in Table 6. Here, we assumed that large innings pitched is most important, small hit ratio is second most important, and small FIP is third most important. As a result, innings pitched is most important and its weight is 0.565.

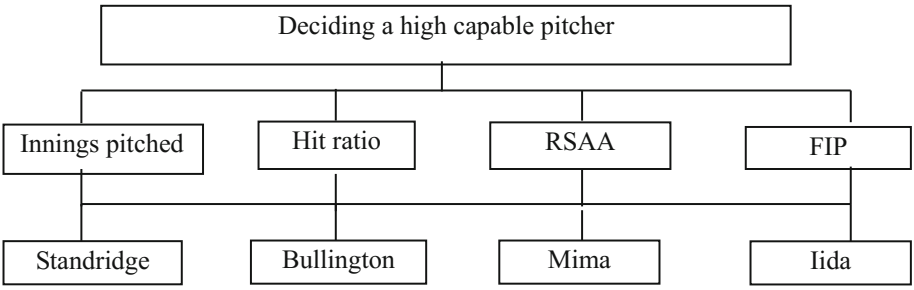


Fig. 6. AHP model for deciding a high capable pitcher.

Table 6. Pairwise comparisons of four criteria.

	Innings pitched	Hit ratio	RSAA	FIP	Weight
Innings pitched	1	3	7	5	0.565
Hit ratio	1/3	1	5	3	0.262
RSAA	1/7	1/5	1	1/3	0.055
FIP	1/5	1/3	3	1	0.118

Consistency index = 0.039

Consistency index shows whether the pairwise comparison is appropriate or not. When the index is lower than 0.1, the pairwise comparison is appropriate. When the index is over 0.1, the comparison is not appropriate and should be corrected. In this case, consistency index was 0.01 and the pairwise comparison was appropriate.

The pairwise comparisons of four alternatives with respect to innings pitched are shown in Table 7. The weight of Standridge was highest, because the number of innings pitched of Standridge was largest.

Table 7. Pairwise comparisons of alternatives with respect to innings pitched.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	6	5	8	0.636
Bullington	1/6	1	1/2	5	0.127
Mima	1/5	2	1	6	0.195
Iida	1/8	1/5	1/6	1	0.042

Consistency index = 0.086

The pairwise comparisons of four alternatives with respect to hit ratio are shown in Table 8. The weight of Iida was highest, because the hit ratio of Iida was smallest.

Table 8. Pairwise comparisons of alternatives with respect to hit ratio.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	1/2	2	1/3	0.154
Bullington	2	1	4	1/2	0.288
Mima	1/2	1/4	1	1/5	0.081
Iida	3	2	5	1	0.477

Consistency index = 0.007

The pairwise comparisons of four alternatives with respect to RSAA are shown in Table 9. The weight of Bullington was highest, because the RSAA of Bullington was largest.

Table 9. Pairwise comparisons of alternatives with respect to RSAA.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	1/5	2	1/2	0.125
Bullington	5	1	6	3	0.577
Mima	1/2	1/6	1	1/3	0.077
Iida	2	1/3	3	1	0.222

Consistency index = 0.011

The pairwise comparisons of four alternatives with respect to FIP are shown in Table 10. The weight of Bullington was highest, because the FIP of Bullington was smallest.

Table 10. Pairwise comparisons of alternatives with respect to FIP.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	1/6	1/2	1/3	0.079
Bullington	6	1	5	2	0.533
Mima	2	1/5	1	1/2	0.130
Iida	3	1/2	2	1	0.253

Consistency index = 0.008

Table 11 shows final results of AHP. Standridge was the most capable pitcher, because we assumed that large innings pitched is most important and small hit ratio is second most important. The number innings pitched of Standridge is largest.

Table 11. Final results of deciding a high capable pitcher.

Criteria	Innings pitched	Hit ratio	RSAA	FIP	Result
Weight of criteria	0.565	0.262	0.055	0.118	
Standridge	0.636	0.154	0.125	0.079	0.416
Bullington	0.127	0.288	0.577	0.533	0.242
Mima	0.195	0.081	0.077	0.130	0.151
Iida	0.042	0.477	0.222	0.253	0.191

Figure 7 shows an example of the relative measurement AHP model created for the task of deciding a reinforced starting pitcher. Here, we used the following five criteria: capability, salary, age, right/left throw and feasibility.

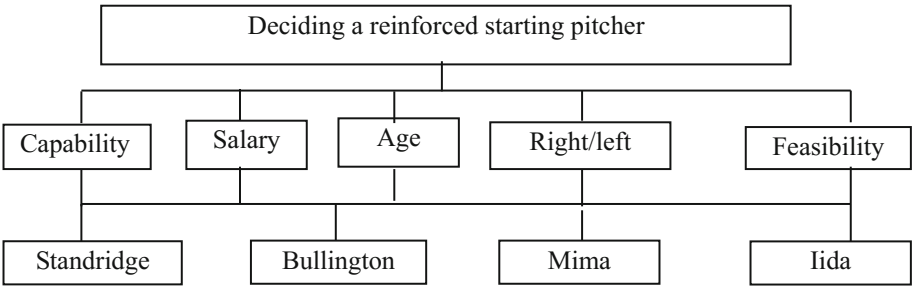


Fig. 7. AHP model for deciding a reinforced starting pitcher.

We assumed the pairwise comparison matrix for Chiba Lotte Marines. The pairwise comparison matrix for the five criteria is shown in Table 12. Here, we assumed that capability and feasibility are most important, and right/left throw is third most important. As a result, capability and feasibility are most important and their weights are 0.362.

Table 12. Pairwise comparisons of five criteria.

	Capability	Salary	Age	Right/left	Feasibility	Weight
Capability	1	7	5	3	1	0.362
Salary	1/7	1	1/3	1/5	1/7	0.039
Age	1/5	3	1	1/3	1/5	0.076
Right/left	1/3	5	3	1	1/3	0.161
Feasibility	1	7	5	3	1	0.362

Consistency index = 0.034

The weights of four alternatives with respect to capability are shown in Table 11.

The pairwise comparisons of four alternatives with respect to salary are shown in Table 13. The weight of Iida was highest, because the salary of Iida was cheapest.

Table 13. Pairwise comparisons of alternatives with respect to salary.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	1/2	1/4	1/6	0.074
Bullington	2	1	1/2	1/4	0.138
Mima	4	2	1	1/2	0.275
Iida	6	4	2	1	0.513

Consistency index = 0.004

The pairwise comparisons of four alternatives with respect to age are shown in Table 14. The weight of Iida was highest, because Iida is youngest.

Table 14. Pairwise comparisons of alternatives with respect to age.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	1/2	1/4	1/6	0.074
Bullington	2	1	1/2	1/4	0.138
Mima	4	2	1	1/2	0.275
Iida	6	4	2	1	0.513

Consistency index = 0.004

The pairwise comparisons of four alternatives with respect to right/left throw are shown in Table 15. The weight of Iida was highest, because left throw is a few and important for Chiba Lotte Marines.

Table 15. Pairwise comparisons of alternatives with respect to right/left.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	1	1	1/2	0.2
Bullington	1	1	1	1/2	0.2
Mima	1	1	1	1/2	0.2
Iida	2	2	2	1	0.4

Consistency index = 0

The pairwise comparisons of four alternatives with respect to feasibilty are shown in Table 16. The weights of Standrige and Bullington were highest, because they declared *free agent*.

Table 16. Pairwise comparisons of alternatives with respect to feasibility.

	Standridge	Bullington	Mima	Iida	Weight
Standridge	1	1	3	3	0.375
Bullington	1	1	3	3	0.375
Mima	1/3	1/3	1	1	0.125
Iida	1/3	1/3	1	1	0.125

Consistency index = 0

Table 17 shows final results of AHP. Standridge was the best. Because we assumed that capability and feasibility are most important. Capability and feasibilty of Standridge are highest. Standridge is selected as the final choice. Actually, Chiba Lotte Marines acquired Standridge as a reinforced starting pitcher.

Table 17. Final results of AHP.

Criteria	Capability	Salary	Age	Right/left	Feasibility	Result
Weight of criteria	0.362	0.039	0.076	0.161	0.362	
Standridge	0.416	0.074	0.074	0.2	0.375	0.327
Bullington	0.242	0.138	0.138	0.2	0.375	0.271
Mima	0.151	0.275	0.275	0.2	0.125	0.164
Iida	0.191	0.513	0.513	0.4	0.125	0.238

5 Conclusion

We proposed a way of clustering professional baseball players with SOMs, considering several team reinforcement strategies using player maps, and deciding team reinforcement strategy with AHP. We used data of pitchers of Japanese professional baseball teams. We used data of pitchers in Japanese baseball teams. First, we collected data of 302 pitchers and clustered these pitchers using fourteen features. Second, we created pitcher maps of all teams and each team with SOM. Third, we examined main features of each cluster. Fourth, we considered team reinforcement strategies by using pitcher maps. Finally, we used AHP to determine the team reinforcement strategy. In future work, we will apply our way to the other sports such as football and basketball. We will use other types of AHP [7] and ANP [12] for decision making.

References

1. Tolbert, B., Trafalis, T.: Predicting major league baseball championship winners through data mining. Athens J. Sports (2016). https://www.athensjournals.gr/sports/2016-3-4-1-Tolbert.pdf
2. Ishii, T.: Using Machine Learning Algorithms to Identify Undervalued Baseball Players (2016). http://cs229.stanford.edu/proj2016/report/Ishii-UsingMachineLearningAlgorithmsToIdentifyUndervaluedBaseballPlayers-report.pdf

3. Pane, M.: Trouble with the Curve: Identifying Clusters of MLB Pitchers using Improved Pitch Classification Techniques (2013). http://repository.cmu.edu/cgi/viewcontent.cgi?article=1184&context=hsshonors
4. Tung, D.: Data Mining Career Batting Performances in Baseball (2012). http://vixra.org/pdf/1205.0104v1.pdf
5. Vazquez Fernandez de Lezeta, M.: Combining Clustering and Time Series for Baseball Forecasting (2014). https://repositorio.uam.es/bitstream/handle/10486/661046/vazquez_fernandez_de_lezeta_miguel_tfg.pdf
6. Kohonen, T.: Self-Organizing Maps. Springer, New York (1995)
7. Saaty, T.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
8. Kohara, K., Tsuda, T.: Creating product maps with self-organizing maps for purchase decision making. Trans. Mach. Learn. Data Min. **3**(2), 51–66 (2010)
9. Doizoe, J., Kohara, K.: Clustering and visualization of goods with self-organizing maps. In: Proceedings of 70th Annual Convention of Information Processing Society of Japan, vol. 4, pp. 911–912 (2008). (in Japanese)
10. NPB (Nippon Professional Baseball Organization). http://npb.jp/
11. Professional Baseball Data. http://baseballdata.jp/
12. Saaty, T.: The Analytic Network Process. Expert Choice, Arlington (1996)