A thesis submitted in partial satisfaction of the requirements
for the degree of Master of Computer Science and Engineering
in the Graduate School of the University of Aizu

Classification and Clustering of Soccer Data

by

Akitaka Moriyama

March 2018

The thesis titled

Classification and Clustering of Soccer Data

by

Akitaka Moriyama

is reviewed and approved by:

Main referee

Associate Professor

Maxim Mozgovoy

Professor

Vitaly Klyuev

Professor

Ihor Lubashevsky

THE UNIVERSITY OF AIZU

March 2018

Contents

Chapter 1 Introduction **1**
1.1 Background 1
1.2 Objective 2

Chapter 2 Related works **3**

Chapter 3 Dataset **4**
3.1 TRACAB 4
3.2 Data format 5
3.3 Preparation for processing 6

Chapter 4 Method **7**
4.1 Clustering 7
4.2 Measurement of similarity between situations 8
4.3 Method for assessment of our similarity funcitons 10
4.3.1 ROC curve and AUC 10
4.3.2 Experiment 12

Chapter 5 Result and Discussion **14**
5.1 Assessment of our similarity functions 14
5.2 Clustering with new similarity function 15

Chapter 6 Conclusion **19**

List of Figures

Figure 4.1 Correspondence between the players 10
Figure 4.2 Confusion matrix 11
Figure 4.3 Experimental scenery 13

Figure 5.1 Query situation 14
Figure 5.2 Previous similarity function 15
Figure 5.3 New similarity function 15
Figure 5.4 ROC curve of previous similarity function. AUC = 0.571 15
Figure 5.5 ROC curve of new similarity function. AUC = 0.777 15
Figure 5.6 Clusters with typical game situations 16
Figure 5.7 Clusters with rare game situations 17

List of Tables

Table 3.1 The data we analyze 4
Table 3.2 Description for the chunk 2 in real soccer game data 5
Table 3.3 Description for the chunk 3 in real soccer game data 5

Acknowledgment

Professor Maxim Mozgovoy provided the soccer simulator to visualize soccer game situations. Yuki Kobayashi, Yuki Konno, Watanabe, and Shoya Kunigita participated our experiment. I would like to thank them and everyone in my laboratory.

Abstract

In this thesis, a method which classifies soccer game situations is proposed. At the present time, sport analytics is a rapidly developed topic, however, the problem of analyzing game strategies in team games, such as soccer, is still difficult to tackle. We believe that it is possible to improve understanding of team strategies and trends in a soccer game through clustering. Game situations of a team in a soccer game can be classified by similar situations. As a data for clustering, real soccer game recording data is used. A similarity function relied on ball/player distance is adopted to measure similarity between situations for clustering using the data. Next, we make sure the similarity function between situations is better than the similarity function used in our previous research. ROC (receiver operation characteristics) curve and AUC (area under the curve) are calculated to assess our similarity functions. As a result, the new similarity function is exhibited better than the previous similarity function. Here, clustering analysis was performed using the new similarity function. Clustering continued as long as the similarity between each cluster was less than 1. As a consequence of clustering, the total number of clusters became 48, and consideration of the team trends are conducted. Also, influence of the new similarity function in clustering is observed.

Chapter 1

Introduction

1.1 Background

Data analysis is the existence that we cannot separate from our life. In modern times various companies adopt data analysis as management strategies, and conduct good efficiency of management and get trends in the market. Our daily life becomes more comfortable due to data analysis in information technology.

Data analysis is also used in the field of sports. For example, ”Moneyball”, the popular book written by Michael Lewis in 2003, tells how general managers built a team on a limited budget by valuing statistics over instinct [1]. Also, sports analysis is not only for players, but also for coaches, spectators, and television viewers. The analysis results are shown in real time when sport events are broadcast live on television. By looking at the statistics of games, we can know invisible information and how to enjoy games further. In the field of sports, data collection and analysis are attracting a great deal of attention for strengthening athletes and teams as well as new elements for entertaining fans.

Soccer game is not an exception. However, prediction of group games of player action and trends of the team such as soccer game is not easy. Skills and strategies are advancing as time goes on, so detailed analysis of team tactics becomes and important problem in sport analytics. In addition, there are important issue, it is the players individual abilities. These are as important in soccer as in other sport games (such as baseball), but they are more difficult to measure. For example, baseball players are usually engaged in well-defined actions such as throw, hit, catch, error, and run, it is easier to measure their performance. A soccer player needs a good combination of skills, which makes analysis more difficult.

1.2 Objective

Understanding team trends and strategies is expected by classifying soccer game situation through clustering proposed in this thesis. Trends of a team from a cluster with many situations and few situations are aware. Also, there is a room of discussion to improve our function of comparing situations. In the process of classification, it is necessary to digitize how similar it is when comparing situations. Comparing function is proposed in our previous research [2], however, the function is not concerned the ball position. There are possibilities that two situations are classified as the same when distance assigned players is small and balls are distant in situation comparison of a team. In this thesis, it is not considered as similar situations. The most important factor in a soccer game situation is assumed players close to the ball. Finally, whether new function was better than the previous function was assessed.

Chapter 2

Related works

Analysis in soccer games is popular, there are researches such as player tracking [3] and evenet detection [4] [5]. For analysis of trends and strategies, there are many analysis methods based on actual images, such as [6].

This is an example of analysis using real soccer game data recording [7], because the paper do clustering using dataset which added event data such as out for goal kick and out for corner, we consider that it cannot apply about analysis of around events.

This paper uses real soccer recording provided by Data Stadium .Inc [8]. In this thesis, similarity is calculated by comparing situations, however, the paper similarity is calculated by comparing pair of pass sequences. This becomes possible to detect detail attack like side-attack and center-attack.

There are other researches like [9] which finds interesting path patterns. However, there are not many papers on analysis of trends and tactics. This is due to difficulties of availability of real soccer game recording data.

We have actual soccer game record data. The data used in this paper is game data of the J1 league. The J1 League is the best league in Japan. In this thesis, as an analysis example of using real soccer data recording, understanding trends and strategies by a method which classify soccer game situations are introduced.

Chapter 3

Dataset

3.1 TRACAB

TRACAB is a tracking system for real soccer game invented by Chyronhego in Sweden [10]. By using exclusive cameras and software, the system tracks players, ball, and referee, and make into data from soccer game video in real time. This system adopts SAAB's proprienary military technology to process data [11]. This TRACAB technology has track record which adopted in Premier League, Bundesliga, and FIFA World Cup as an official data tracking system. Today, the system is used in J1 League game match which is the best League in Japan [12].

The dataset used in this thesis is created by this system and given by Data Stadium Inc. [13]. We have 5 games and 6 teams dataset. The game data are Koube versus Nagoya, Urawa versus Yokohama, Nagoya versus Shimizu, Shimizu versus Yamagata, and Yokohama versus Koube. The detail description is table 3.1

Table 3.1: The data we analyze

Game	Data
Koube versus Nagoya	July 9, 2011
Urawa versus Yokohama	May 3, 2011
Nagoya versus Shimizu	May 7, 2011
Shimizu versus Yamagata	June 15, 2011
Yokohama versus Koube	June 5, 2011

Table 3.2: Description for the chunk 2 in real soccer game data

Properties	Valid values	Remarks
Team	0, 1, or 3	”1=Home team, 0=Away team, 3=Referee Other values are used for internal purposes.
System target ID	1 to 29	
Jersey Number	-1 or 1 to 99	Jersey numbers are 1 - 99. Jersey -1 is ”unassigned” except for team 3 (the referee)
Position X	-5250 to 5250	
Position Y	-3400 to 3400	
Speed	0 to infinity	

Table 3.3: Description for the chunk 3 in real soccer game data

Position X	-5250 to 5250	
Position Y	-3400 to 3400	
Position Z	0 to infinity	Note that the position of the ball center is displayed. Normal Z position for a ball on ground is thereby 10 cm.
Speed	0 to infinity	
Ball owning Team	’H’ or ’A’	”H’=Home team, ’A’=Away team
Ball status	”Alive” or ”Dead”	”Alive”=In play, ”Dead”=Not in play

3.2 Data format

The dataset of real soccer game recording used in this thesis consists of colon-separated chunks. They can be Integer or String chunks. Strings can be just strings or they can be arrays of objects represented as strings. In the latter case the positions in the array are separated with semicolons. The individual properties of each object are separated with commas. The data set of real soccer recordings consists of three main chunks. The first is an integer chunk containing the frame count of the current frame. The frame count is unique for each frame, and is generated by the tracking system. The second chunk is an array of 29 player and referee candidate target objects. The last chunk is an array of one object: the ball. The chunk 1 is just integer data. The data type of chunk 2 is string-represented array of up to 29 objects. Each object contains the properties shown in 3.2. The data type of chunk 3 is a string-represented array of one object. This object contains the properties shown in 3.3

3.3 Preparation for processing

Sometimes soccer games are suspended. We assume that team formations in game pauses is not important. In other words, for example, when the ball is out of field bounds, the ball crosses the goal line, or the referee stops the game due to a foul. The teams do not make formations during the game pause until they resume play, so we only extract situations from the non-suspended fragments of the game.

Chapter 4

Method

To classify soccer game situations, measurement of similarity between situations are required. We describe 2 similarity functions and ascertain the quality by using ROC curve. Finally, we classify soccer game situations by using the similarity function.

4.1 Clustering

Clustering is a method for machine learning, which needs input data, and there is no given correct answer. It is used to discover and predict new characteristics of data. For the purpose of the present research we decided to adopt a hierarchical clustering procedure, since it allows further analysis of individual cluster structure, which helps to make more fine-grained conclusions about analyzed games. Hierarchical clustering algorithms are divided into two classes: divisive and agglomerative [14]. In this paper, we adopt a popular and simple agglomerative clustering procedure. It is performed in a bottom-up way. We assume that a similarity between clusters is 0 to 1 (0 is similar, 1 is dissimilar). We set each game element object as a separate cluster, then each pair of closest clusters merged recursively until a similarity between clusters is less than 1. Our algorithm consists of the following steps.

Step 1: We calculate the similarities between all the clusters in the dataset. We store the calculated similarities in a distance matrix, and proceed to the Step 2.

Step 2: We check condition of termination. When we finished to do clustering, there is a possibility of all similarities between each cluster become 1. We then finish clustering. Otherwise we go to the Step 3.

Step 3: We find a pair of a clusters to be merged together. By using the distance matrix, we can search for the minimal similarity between previously unused clusters. Next, we proceed to the Step 4.

Step 4: We merge the pair of clusters to make a new cluster. Since the obtained cluster needs a centroid, we set as a centroid of the combined pair the cluster containing more game situations. If the number of containing situations of obtained cluster is the same, we set as a centroid of the pair the cluster which is the close to the ball. After this operation we return to the Step 2.

Here, we need to consider how the similarity between clusters calculate. In next section, similarity function is mentioned to measure a similarity between clusters.

4.2 Measurement of similarity between situations

The problem of game situations similarity calculation can be regarded as an assignment problem, which is a classical task of mathematical programming and optimization [15]. In assignment problem, the task is to find the optimal correspondence between the elements of two sets. Each correspondence incurs a certain cost, so the goal is to minimize the total cost. For example, suppose there are three workers and three jobs. One worker can take strictly one job, and there should be no role duplication. The time required to finish each job is different for different workers, so we need to minimize the total time. In our case, the goal can be states which is the optimal assignment of player pairs across two different game situations to minimize the total distance between the players. Let us call S and T a certain situation. $S = \{1, ..., 11\}$ and $T = \{1, ..., 11\}$ are given. In this problem, the cost when s is assigned to t is c_{st}, $c_{st} = D(s, t)$ in $s \in S$, $t \in T$. $D(s, t)$ is a distance between s and t; s and t are a player of a team of each situation. In addition, 0-1 variable x_{st} is prepared which is given $x_{st} = 1$ when s is assigned to t or $x_{st} = 0$ when it is otherwise. Then, the formulation of this problem is below:

minimize cost

$$= \sum_{s=1}^{11} \sum_{t=1}^{11} c_{st} x_{st}, \qquad \forall s \in S, \forall t \in T$$

subject to

$$\sum_{t=1}^{11} x_{st} = 1, \qquad \forall s \in S$$

$$\sum_{s=1}^{11} x_{st} = 1 \qquad \forall t \in T$$

Here, we consider about c_{st}, or $D(s,t)$. In previous research, $D(s,t)$ was defined as Euclidean distance [2]. Regarding a player of s, the x-coordinate and y-coordinate is represented as s_x and s_y. The same can be said to t. Then, the formulation is $D(s,t) = \sqrt{(s_x - t_x)^2 + (s_y - t_y)^2}$. We present the similarity function of the previous research as below to present the similarity from 0 to 1.

$$c_{st} = Sim(s,t) = 1 - \frac{1}{D(s,t)+1}$$

For Hungarian algorithm, a similarity needs to be high degree when the cost is small. Therefore, we defined as $1 - \frac{1}{D(s,t)+1}$ for convenience. The formulation means that the cost is close to 0, the set of situations is similar, while the cost close to 1, it is not similar. In this thesis, a new similarity function is proposed. We consider that the most important place in the pitch of soccer game situation is close to the ball. Then, we affect the Euclidean distance relied on player and ball to the similarity. A Euclidean distance between the place of the mean coordinates of measuring 2 players u and the place of the ball b within a certain range is called as weight w and divide the similarity by the weight. The range is set as 1000, it is about 20 meters long in real soccer game pitches. Also, weight is varied from 1 to 10. The Euclidean distance between the mean place of players and the ball is close, the weight becomes 10, otherwise, it becomes 1. The weight makes 1 when the Euclidean distance between the mean place of players and the ball over the range. The formulation is as below.

$$c_{st} = \frac{Sim(s,t)}{D(u,b)} = \frac{1 - \frac{1}{D(s,t)+1}}{D(u,b)}$$

Finally, a similarity between situations is the minimum of the sum of each cost, we divide the similarity by 11 in order to present it from 0 to 1. Here, demonstration the new similarity function is better than the previous one is needed. Hence, the method for assessing similarity functions is explained in the next chapter.

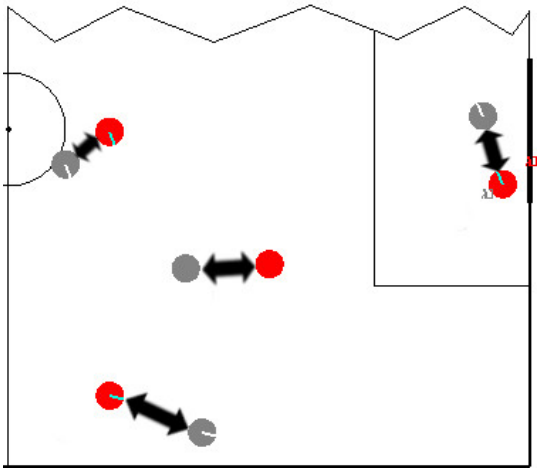

Figure 4.1: Correspondence between the players

4.3 Method for assessment of our similarity funcitons

4.3.1 ROC curve and AUC

For assessment of our two approaches which we mentioned in previous section, we present a graph of ROC curve (receiver operating characteristic curve) [16]. This is that illustrates the classifying ability of a binary classifier system as its discrimination threshold is varied. It was originally developed as a communication engineering theory of the radar system and was developed as a method for detecting the existence of enemy aircraft from the noise of the radar signal. In clinical research, it is often used as a method to evaluate the usefulness of diagnostic tests as a method of evaluating the strength of the relationship between independent variables as continuous variables and outcomes as binary variables. ROC analysis has been used in medicine, radiology, biometrics, model performance assessment, and other areas for many decades and is increasingly used in machine learning and data mining research.

We first consider a binary classification. We can get a similarity between situations by using similarity function. When a threshold is given, there are two outcomes by the similarity. Each of outcomes are able to be labeled as positive (p) or negative (n), an outcome regards as (p) if the similarity is bigger than the threshold, then total outcomes are four. If the outcome from a prediction is (p) and the actual value is also (p), then it is called a true positive (TP). However, if the actual value is (n), it is a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are (n), and false negative (FN) is when the prediction outcome is (n) while the actual value is (p). These outcomes by calculating

similarities between situations when a threshold was given can be expressed as confusion matrix like 4.2.

		Predicted Condition	
		P (positive)	N (negative)
True Condition	P	TP (True Positve)	FN (False Negative)
	N	FP (False Positive)	TN (True Negative)

Figure 4.2: Confusion matrix

Then, we can calculate true positive rate and false positive rate as below

True Positive Rate

$$TPR = TP/(TP + FN)$$

False Positive Rate

$$FPR = FP/(FP + TN)$$

TPR (true positive rate) is the rate that positive of true class predict positive correctly. It is also called as sensitivity. FPR (false positive rate) is the rate that positive of true class predict positive incorrectly and known as $1 - specificity$. $1 - specificity$ has the same meaning of FPR because $specificity$ is presented as $TN/(FP + TN)$. Since the two rates are calculated based on the total number of data in each correct positive condition data, there are no effect of bias of the number of data between conditions. Then, we vary a threshold at a certain rate and plot the TPR and FPR for each threshold on the vertical and horizontal axes respectively to illustrate the ROC curve. By using ROC curve, we can assess our methods by AUC (area under the curve). AUC is the area under the ROC curve and indicates the goodness of the ability of the similarity functions. The AUC close to 1 is a classifier which can classify correctly, and when a classifier predicts randomly, the AUC becomes 0.5. Thus, the similarity function of AUC which is high area can classify better accurately than the low area one.

4.3.2 Experiment

To classify soccer game situations, measurement of similarity between situations are required. We describe 2 similarity functions and ascertain the quality by using ROC curve AUC. At first, we need to create a dataset in which certain situations are correctly classified whether similar or not for a query situation. The chosen dataset to assess our methods is Nagoya versus Shimizu game. There are 44934 frames (situations) in the 1st half, and 35628 frames in the 2nd half. We have extracted 296 frames from each half of the dataset, and processing for the total of 538 frames was performed. Moreover, there is a home team and an away team in a frame. In order to compare both teams in a frame, we compared the total of 1076 situations in practice by moving the team on the right-side point symmetrically around the pitch. A query situation is chosen from the extracted situation randomly. This experiment is conducted using the data of the same extracted situations and the query situation for all research participants.
The number of people who participated in our experiment is 4. The experimental scenery is like the picture below 4.3. The bottom screen shows the query situation, and the top screen shows situations which were extracted from Nagoya versus Shimizu game.

After the extraction of situations from Nagoya versus Shimizu, a similarity between the query situation and an extracted situation by a similarity function is calculated, and both the similarity and the situation data are randomly outputted to a file. Next, my algorithm read situation data from the outputted file, and display a picture of the situation at top screen. The query situation is displayed at the bottom screen. Then, research participants compare the top and bottom situation and push a button of ”similar” or ”not similar”. My algorithm receives 0 when ”similar” button is pushed, otherwise, it receives 1. This experiment is finished after research participants compare 1076 situations. Finally, the ROC curve is drawn by using the input and similarities which were calculated by a similarity functions. At this time, the range of threshold that can take is from the minimum similarity to the maximum similarity of the outputs of the similarity functions. The amount of increase of the threshold is obtained by dividing the value calculated by subtracting the minimum value of the similarity from the maximum value of the similarity by 100. Each ROC curve is displayed by input value and similarity by each similarity function. Also, we determine whether similarity function better by calculating the AUC of each ROC curve.

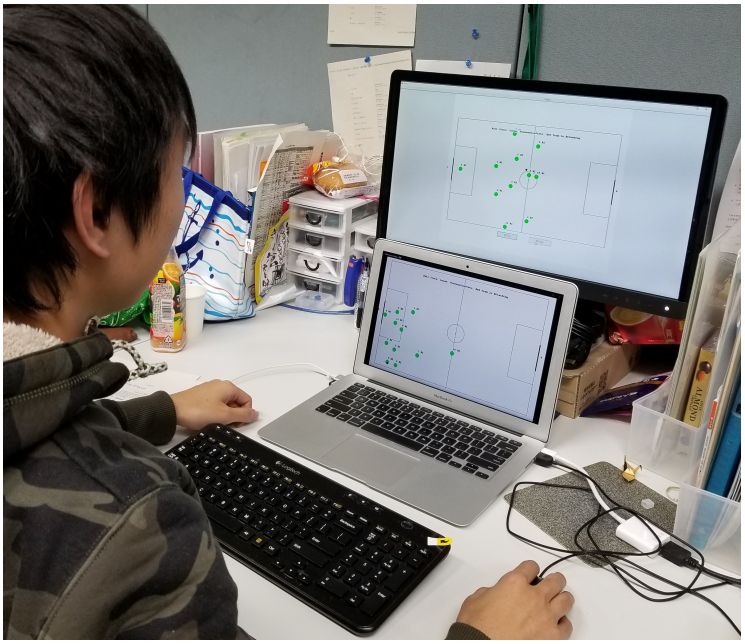

Figure 4.3: Experimental scenery

Chapter 5

Result and Discussion

5.1 Assessment of our similarity functions

The state of the chosen query situation from Nagoya versus Shimizu game is that most players in the team side gathered with the left half of the team side 5.1. Two players are isolated from a group of 9 players. We checked the situation at this moment of this real soccer game on video, the situation was that the team defended from the opposite team. The ball is placed near the bottom left corner of the pitch, and it was a before scene that try to score a goal by a player of the opponent team who kicking the ball from our side corner.

5.2 and 5.3 are the result of classification.

,

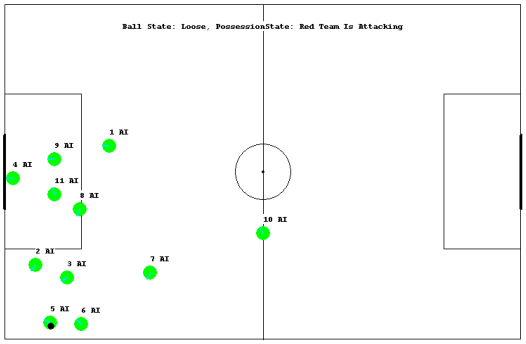

Figure 5.1: Query situation

The most similar situation of each similarity function was omitted because it is the same as the query situation. In the situation of the new similarity function, the player who is distant from the ball has less correspond with the situation of the query. However, around the ball, it seems as if the situation of the query is laying the same formation than the normal situation.

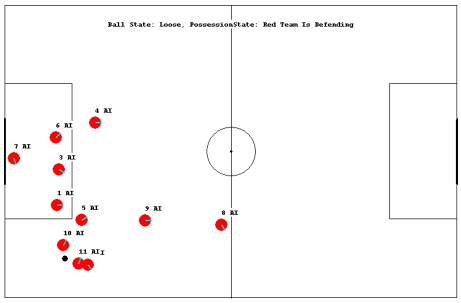

Figure 5.2: Previous similarity function

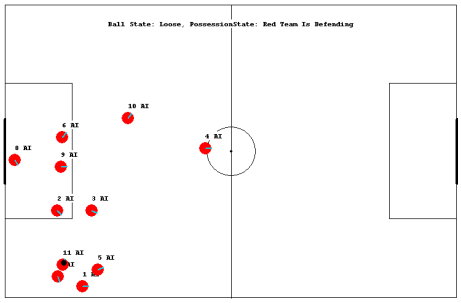

Figure 5.3: New similarity function

We then show the result of ROC curve graph and AUC by 1 in 4 examination participants. Left picture is a situation by using the previous similarity function 5.4, right picture is a situation by using the new similarity function 5.5.

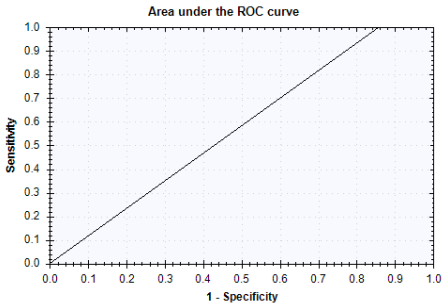

Figure 5.4: ROC curve of previous similarity function.
AUC = 0.571

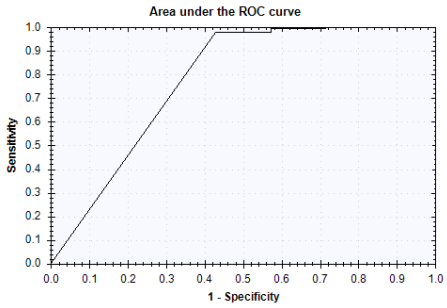

Figure 5.5: ROC curve of new similarity function.
AUC = 0.777

The AUC of the ROC curve by the new similarity function is exhibited better than the previous one. The AUC of the previous similarity function is 0.571, and the new similarity function 0.777. The AUC by the previous similarity function of other 3 examination participants were 0.528, 0.571, 0.506, however, the AUC by the new similarity function was 0.685, 0.703, 0.645. Thus, the new similarity function can be confirmed to be better than the previous one.

5.2 Clustering with new similarity function

The pictures of 5.6 and 5.7 are the result of clustering. Shimizu team of Shimizu versus Yamagata game is set to do clustering and extracted 25 each situation.

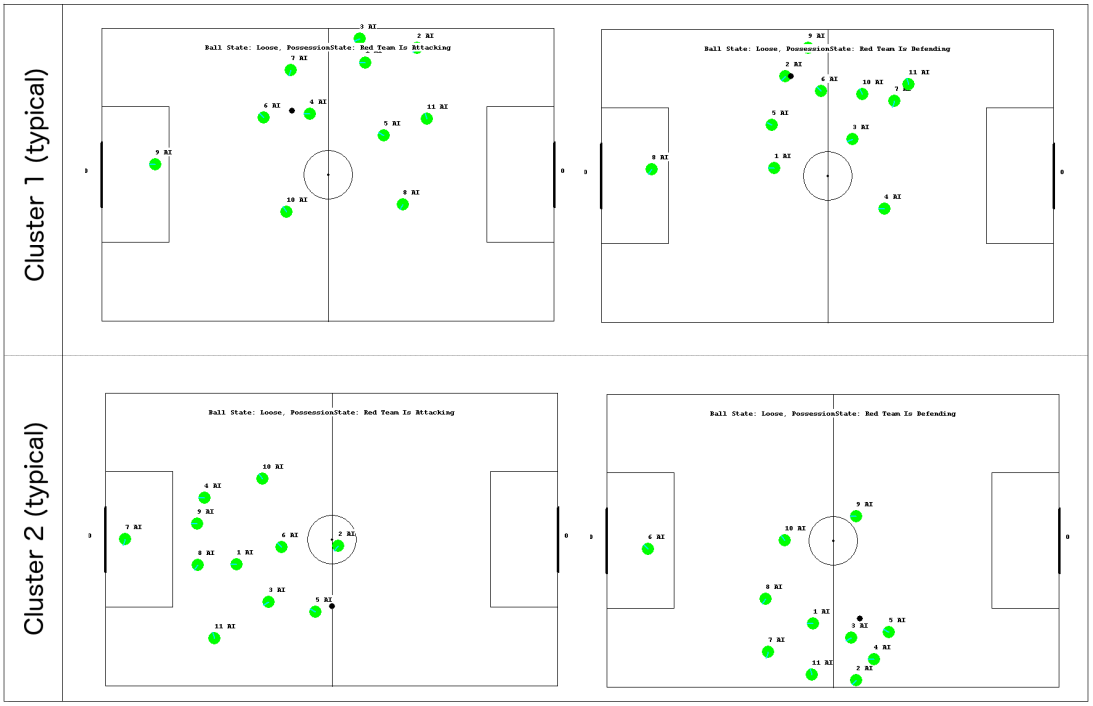

Figure 5.6: Clusters with typical game situations

When the number of situations which belongs to a cluster is large, we call it ”typical situation”. It means that the team did like the formation many times in the game. Otherwise, we call it ”rare situation”. The number of situations which belongs a cluster is small, it means that the situation occurred few times in the game. Analysis of typical situations and rare situations has possibilities of leading interesting outcomes.

In our experiments, we can be classified 48 clusters and the number of situations which was concluded in a cluster by descending order is 187, 172, 155, ... , 5, 3, 2, 1.

The typical situations in the cluster 1, these situations are chosen from a cluster which has 155 situations randomly. In cluster 2 of typical, it has 172 situations. In cluster 1, some are similar situations and others are dissimilar situations like the 2 situations because the number of situations that belongs to a cluster is large. When we watched the video of the moment, the left situation was a scene that defend our goal from opponent team’s sudden attacking. The same thing as the left situation, the right situation was the moment to took the ball and kicked it out to the line from opponent team player’s attacking. We can be considered as attacks at the upper middle of the pitch. Also, in cluster 2, the situations are not similar like in cluster 1, however when we watched the video of the game, the left situation was the scene just before

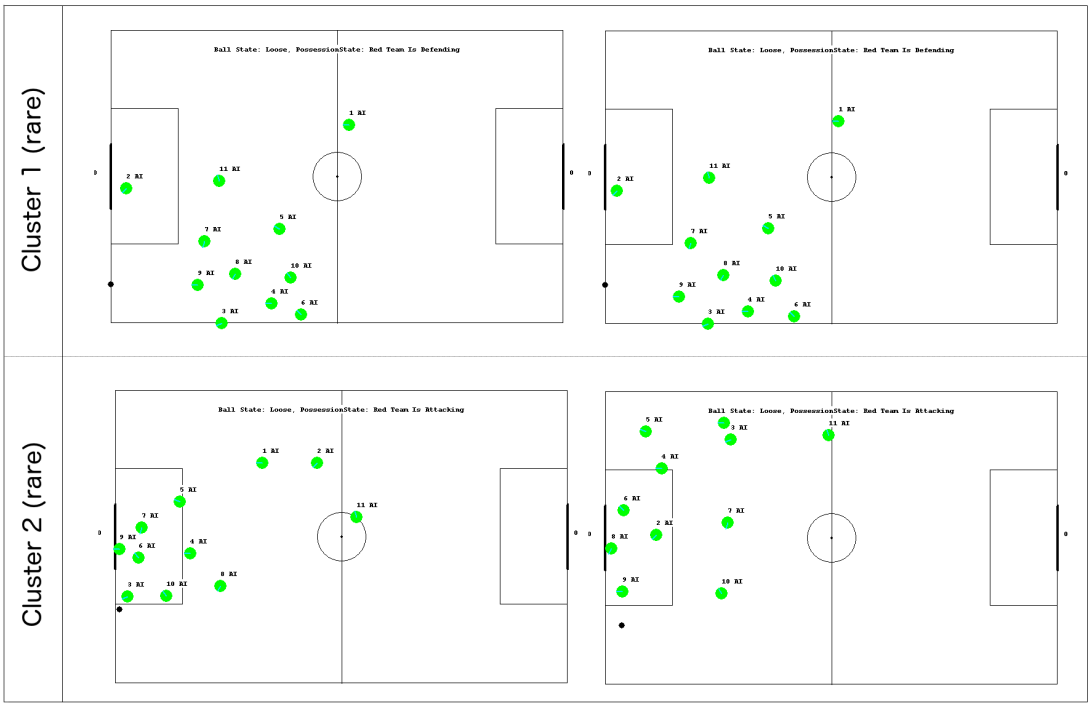

Figure 5.7: Clusters with rare game situations

being attacked, with the ball kicked out towards the upper left corner. We were able to extract situations that are typical in soccer games, such as attacking and defending.

The rare situations in the cluster 1, these are chosen from a cluster which has 5 situations. In cluster 2 of rare, it has 3 situations. Unlike the typical situations, rare situations are similar to each situation in the cluster from the pictures of these situations. In cluster 1, we can extract almost similar situations. In cluster 2, as a result of watched each situation in the video of the game, both situations was the scene that an opponent player kicked the ball from the lower left corner and opponent players in front of the goal tried to score a goal. In this cluster, when we focused of the 4 players in front of the goal, it looks like doing the same formation at each situation. We can be considered that they play a role of defending from opponent team's attacking.

An example situations of affecting new similarity function is rare cluster 2. As we mentioned above, 4 players in front of the goal is similar, however, distant players from the ball is not

similar. In addition, when we watched the situation at the moment by video, distant players from a group of 9 players in front of the goal are not affect to play after these situations. This is influenced that players close to the ball become important factors in soccer games.

Chapter 6

Conclusion

In this paper, we gained understandings of strategies of a team. As a data for clustering, a real soccer game recording data provided by DataStadium Inc was used. The dataset is 25 frames per seconds and a frame consist from 3 chunks: a frame number, players and referee, and the ball. Each player and referee include information like coordinates. We made similarity functions to do clustering by using the dataset. In our past research, we proposed a new measuring method to measure a similarity between 2 clusters (situations), in this thesis the idea which players near the ball have importance about strategies were affected. Next, to make sure that the new similarity function is better than previous one, we examine to measure ROC(receiver operation characteristic) curve and AUC(area under the curve) by 4 examination participants. They judge a query situation and another situation whether similar or not. The game data of Nagoya vs Shimizu was used to examine and compared 1076 situations to create true condition data. By using the similarities generated by a similarity function for each situation of true condition data, ROC curve and AUC are calculated. As a result, the AUC of the new approach is larger than the previous one in all examination participants, and the new approach is better. Here, clustering analysis was performed using new approach. A hierarchical clustering was adopted, and Shimizu team of the game of Shimizu versus Yamagata dataset was used to do clustering. Clustering continued as long as the similarity between each cluster was less than 1. As a consequence of clustering, the total number of clusters became 48, we were able to get typical situations which the cluster has many situations and rare situations which the cluster has few situations. Attacking and defending situations were obtained from a cluster which has typical situations. In addition, from a cluster which has rare situations, situations like defending from opponent team attacking near the left bottom corner were obtained. Also, we can confirm

that the new similarity function affects to clustering.

We hope that this research will lead to application of creation for data-driven AI (artificial intelligence) agent which is controlled by computer in the future. For example, we consider about a match of human vs agent. By using the similarity function we proposed in this thesis, the most similar situation can be found from a dataset. At this time, all agents can move like a real soccer team player. Thus, we consider that similarity function can be contributed in game AI. Also, we expect that this applies not only to 11-to-11 sports like soccer game but also basketball to games of sports.

References

[1] M. Lewis, *Moneyball: The art of winning an unfair game*. New York: Norton, 2003.

[2] M. M. A. Moriyama, *Classification and Clustering in Soccer Analytics*, 2016.

[3] Y. Ohno, J. Miura, and Y. Shirai, "Tracking players and estimation of the 3d position of a ball in soccer games," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 1, 2000, pp. 145–148 vol.1.

[4] M. Tavassolipour, M. Karimian, and S. Kasaei, "Event detection and summarization in soccer videos using bayesian network and copula," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 2, pp. 291–304, Feb 2014.

[5] Y. Yang, S. C. Chen, and M. L. Shyu, "Temporal multiple correspondence analysis for big data mining in soccer videos," in *2015 IEEE International Conference on Multimedia Big Data*, April 2015, pp. 64–71.

[6] W. Puchun, "The application of data mining algorithm based on association rules in the analysis of football tactics," in *2016 International Conference on Robots Intelligent System (ICRIS)*, Aug 2016, pp. 418–421.

[7] X. Wei, L. Sha, P. Lucey, S. Morgan, and S. Sridharan, "Large-scale analysis of formations in soccer," in *2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov 2013, pp. 1–8.

[8] S. Hirano and S. Tsumoto, "Grouping of soccer game records by multiscale comparison technique and rough clustering," in *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, Nov 2005, pp. 6 pp.–.

[9] H. Hibi, Y. Kameoka, S. Uchiyama, and Y. Yamamoto, "Visualization of attack and analysis of play affect goal in soccer game," in *2015 13th International Conference on ICT and Knowledge Engineering (ICT Knowledge Engineering 2015)*, Nov 2015, pp. 15–18.

[10] "Tracab," http://chyronhego.com/sports-data/tracab, accessed: 2017-12-18.

[11] "Saab," http://saabgroup.com/, accessed: 2017-12-18.

[12] "J league," http://www.jleague.jp, accessed: 2017-12-18.

[13] "Data stadium inc." https://www.datastadium.co.jp, accessed: 2017-12-18.

[14] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[15] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[16] J. A. Hanley and B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, 1982.