Classification and Clustering of Soccer Game Situations

Akitaka Moriyama s1200110 Supervised by Maxim Mozgovoy

Abstract

In this paper, a method for analysis of a soccer team behavior by classification and clustering of soccer game situations is suggested. Analysing the field of sports has become popular in recent years. However, soccer games have complexity which other sports do not have. We calculate the similarity between two situations in a soccer game and then classify the situations on the basis of similarity. As a result, by considering all the clusters we can extrapolate strategies and trends of typical and rare situations.

1 Introduction

Prediction and analysis are often used in the field of sports. The analysis results are shown in real time when sport events are broadcast live on television. Details of analysis are also shown on the Internet [1]. Sports analysis is not only for players, but also for coaches, spectator, and television viewers.

In the soccer game, typical TV analysis shows statistical data, such as the number of free kicks, penalty kicks, offsides, the list of goal scorers, and players' total on-field time. Game overview and team strategy prediction is also often discussed.

However, the analysis of soccer games is not as developed as other sports [2]. The cause is the complexity of soccer. One of the factors is that there are 22 players in the field. Each player is able to run and walk everywhere in the field. So we have to follow all the players at the same time when we analyze the situation. Consequently, a large number of calculations are needed. Another factor 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, and run, it is easier to measure their performance. A soccer player needs a good combination of skills, which makes analysis more difficult.

In this paper, a method which classifies soccer game situations is proposed. By developing the method, we expect to gain understanding of trends and strategies of typical and rare situations of the team. This can be accomplished with clustering by using a preprocessed real soccer dataset. As a preliminary step, we will also need to be able to calculate similarity between clusters to do clustering.

2 Method

2.1 Preprocessing

A real soccer game dataset was provided by the DataStadium company [3]. The games we analyze are Kobe versus Nagoya (9.7.2011), Nagoya versus Shimizu (7.5.2011), Shimizu versus Yamagata (15.6.2011), Urawa versus Yokohama (3.5.2011), and Yokohama versus Kobe (5.6.2011). The structure of this dataset consists of three chunks: the present frame number, each player data, and ball data. Each player has team ID, unique player ID, jersey number, field coordinates, and speed. The ball has the coordinates, speed, ID of the owning team, and status. In this paper, we use the data of Kobe versus Nagoya game played on July 9, 2011.

At the preprocessing stage we select only game situatiions where the following conditions are met:

- Ball status is "alive"
- Number of players in the home team is eleven

Sometimes soccer game are suspended. For example, when the ball is out of field, the ball crosses the goal line, and the referee stops the game due to a foul. The teams do not make formations during the game pause until they resume play. Therefore, we only extract the situations from the non-suspended part of the game. Additionally, in a soccer game, there are two teams: home team and away team. We only follow players of home team, remove frames in which the player number is not eleven and extract about 1000 frames equidistantly. We do this because full processing requires a great amount of computational resources. This preprocessed data is then used during clustering.

2.2 Clustering

For the purpose of the present research we decided to adopt a hierarchial clustering procedure, since it allows to analyze individual cluster structure futher and make more fine-grained conclusions about the nature of the game of soccer.

Hierarchical clustering algorithms are divided into two classes: divisive and agglomerative [4]. In this paper, we adopt a popular and simpler agglomerative clustreing. It is performed in a bottom-up way. We set each game element object as a single cluster, then we merge each pair of closest clusters recursively until a single cluster has all the elements of our dataset. Clusters have its own index in the order of making to merge.

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

Step 2 is the condition of termination. Since we use agglomerative clustreing, we should finish the procedure when a single cluster includes all situations. Otherwise, we go to step 3.

Step 3 is to find a pair of a clusters which will be merged together. By using the distance matrix we can search for the minimal distance between previously unused clusters. Next, we go to step 4.

Step 4 is to merge the pair of clusters and make a new cluster. The new cluster needs to be set a centroid. This time, we set the most influential either pair of clusters which include the most situations. When clusters to merge have the same number of situations, either the highest index cluster becomes a centroid of a new cluster. We then allow the status of the pair of clusters to be used. we then go to step 2.

For this algorithm to work we need to develop a distance function used to compare soccer game situations.

2.3 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. The task is to choose the best combination.

For example, there are three workers and three jobs. The time required to finish each job is different for different workers, so the task is to minimize the total time. One worker can take one job, and there should be no role duplication. The assignment problem is to find the optimal combination. In our case, it can be regarded as the optimal assignment of each player across two situations as in the example of the workers just given.

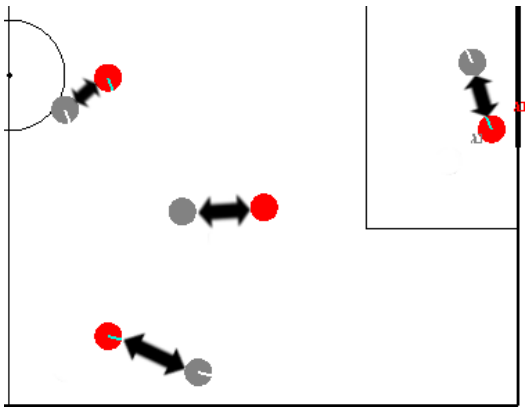

Figure 1: correspondence of players

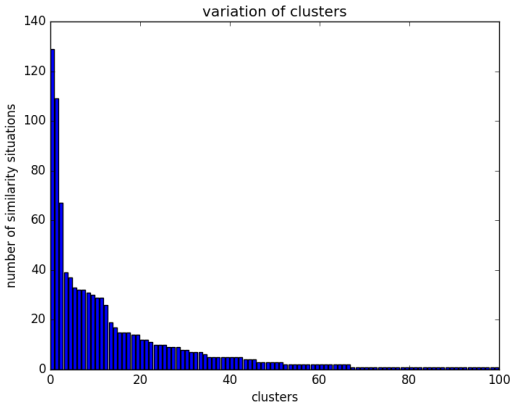

Figure 2: Distribution of game situations in clusters

In this work, we solve the assignment problem with a Hungarian algorithm [5], also known as Kuhn-Munkres algorithm. Its time complexity is $O(N^3)$.

A similarity between two situations is defined as the sum of all Euclidean distances between a pair of assigned players (see Figure 1).

3 Results

The figure 2 shows the histogram of the results of classification. The number of clusters is 100. The vertical axis is the number of situations, contained in each cluster.

Similarity between game situations decrease rapidly with each new cluster. The largest 14 clusters include

about 64 of all game situations. We will presume that the top 14 clusters contain typical game situations of soccer, while the others contain rare game situations.

The Figure 3 shows four typical soccer game situations, taken from two largest clusters identified. Similarly, the Figure 4. shows four rare game situations obtained from the two smallest clusters. The home team is on the right side of the field, the away team is on the left side.

4 Discussion

Almost all the players of the home team is on the field of the away team in the cluster 1 of figure 3. The ball is also in the away team side. Here the home team attacks the away team. The home team makes a formation where the players form vertical lines in the cluster 2 of the figure 3. Many players of the away team are located on the field of the home team, so the home team is attacked by the away team. Thus, the home team makes the formation to defend attacks of the away team.

Both team players are located under the field center in the cluster 1 of the figure 4. Distances between the players is small, so the players scramble for the best position to get the ball. According to the soccer game video, this is a moment which the game resumed play. The home team defend an attack of the away team in the cluster 2 of figure 4. Players make formation where the players form a vertical line. This is a moment which the game resumed play because players of the away team is in the middle of the field.

It is evident from the results that there are differences in formations of each cluster. The team make an optimal formation in each situation. We can observe situations occurring frequently in typical clusters. For example, there are the following frequent cases: players are in the middle of the field and the home team attacks/defends like in cluster1 and cluster2 of the figure 3. There are also numerous situations of the game resumed play in rare clusters like shown in the figure 4.

5 Conclusion

We classified soccer game situations into typical and rare using clustering. In this paper, we defined similarity as the sum of all Euclidean distances between a pair of assigned players. We consider adopting other similarity like variance to improve the accuracy. We hope that this method will contribute to the methods of deeper analysis of soccer games, since it helps to understand the nature of soccer and identify frequent and rare game situations.

Acknowledgement

Professor Maxim Mozgovoy provided the soccer simulator to visualize soccer game situations. Victor Khaustov improved the converter to convert the dataset which the soccer simulator can handle. I would like to thank them and everyone in my laboratory.

References

[1] “J リーグ (J League).” `http://www.jleague.jp`. Accessed: 2016-01-28.

[2] “@IT そもそもサッカーは、データ分析に向くスポーツか (Does soccer game cut out for data analysis?).” `http://www.atmarkit.co.jp/ait/articles/1509/29/news029.html`. Accessed: 2016-01-25.

[3] “データスタジアム株式会社 (Data Stadium Inc.).” `https://www.datastadium.co.jp`. Accessed: 2016-01-31.

[4] A.K. Jain and R.C. Dubes, Algorithms for clustering data, Prentice-Hall, Inc., 1988.

[5] 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.

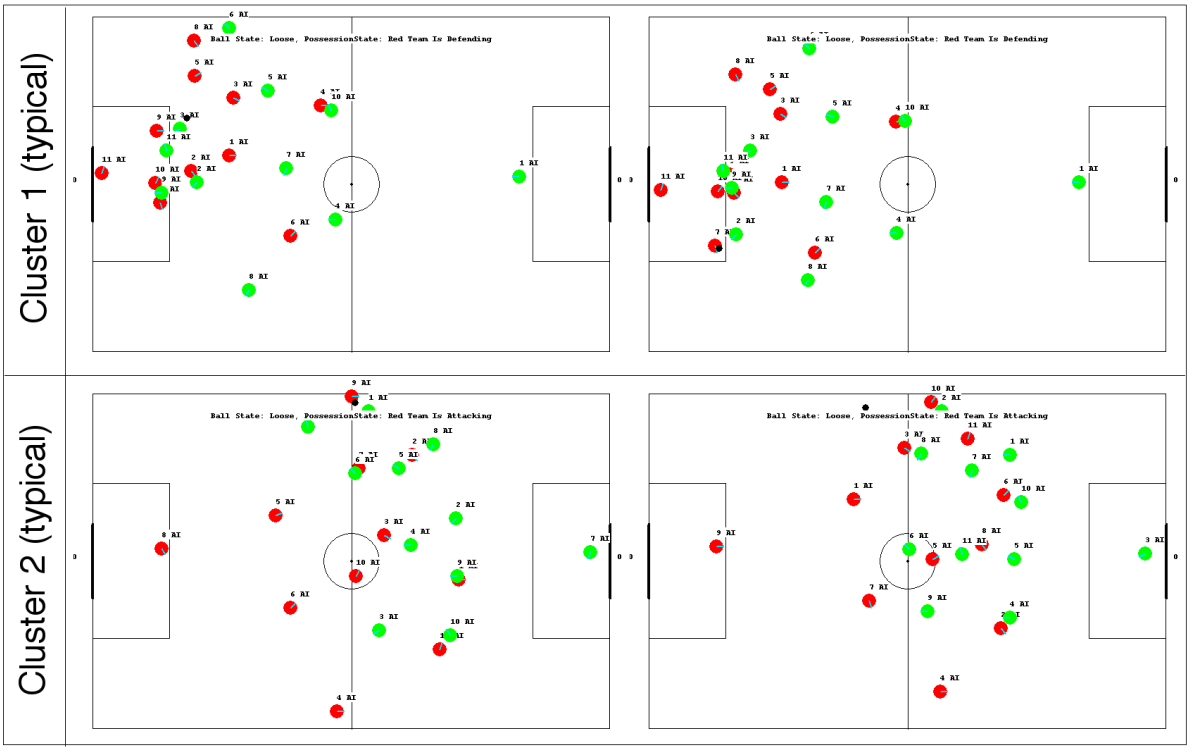

Figure 3: Clusters with typical game situations

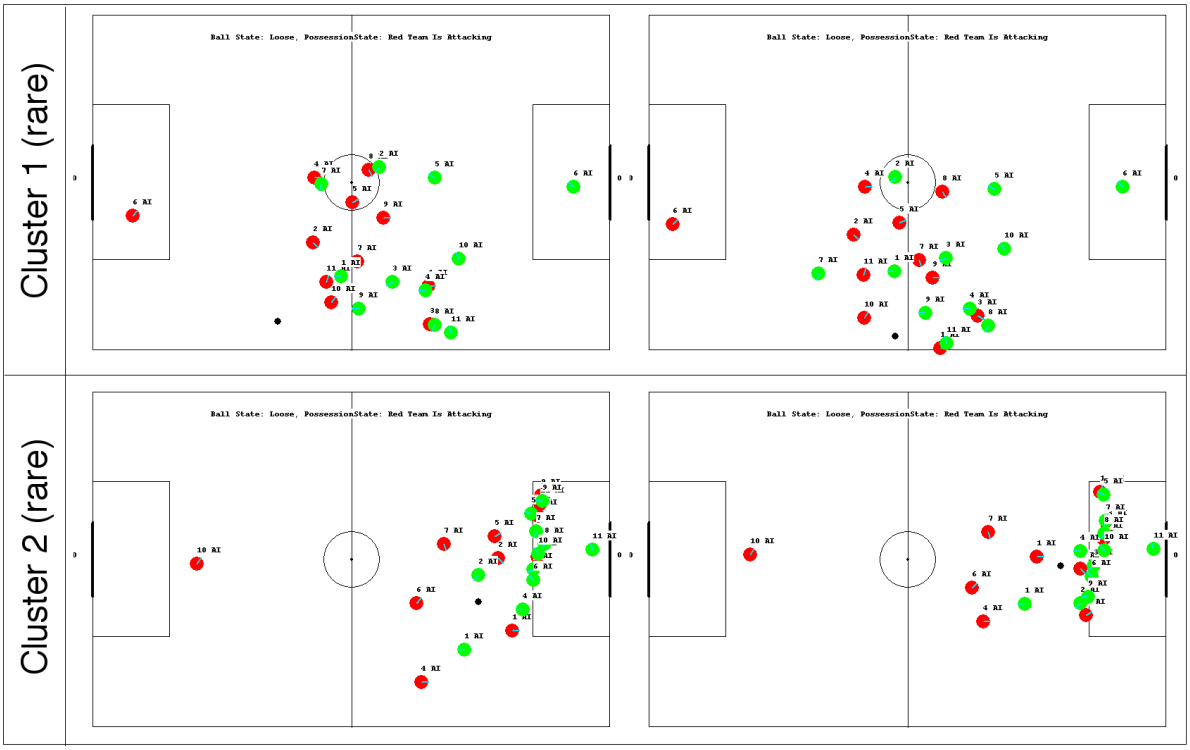

Figure 4: Clusters with rare game situations