ISSN 1750-9823 (print)
International Journal of Sports Science and Engineering
Vol. 07 (2013) No. 02, pp. 067-078

Applying Data Mining to Analyze the Different Styles of Offense between Manchester United and FC Barcelona in the European Champions League Final

Tianbiao Liu [+], Andreas Hohmann
Institute of Sports Science, University of Bayreuth, 95447 Bayreuth, Germany

(*Received June 4, 2012, accepted September 9, 2012*)

Abstract. Soccer games are always analyzed by means of traditional statistical methods. With the development of modern IT technology, new methods were introduced to analyze soccer games. In this study of the European Champions league final between FC Barcelona and Manchester United, detailed game actions were recorded and data mining methods were applied. Microsoft SQL Server Data mining add-ins served to calculate association rules and dependency networks. The data mining results led to typical playing patterns of both teams.

Keywords. Soccer, European Champions league final, Data mining, Association rules, Dependency networks

1. Introduction

In soccer games, offense plays a key role in regard to scoring goals and winning the game. To analyze the different attacking styles of soccer teams, researchers have developed various methods that are based on traditional statistical procedures [1],[2]. In recent years, researchers introduced advanced mathematical models for soccer game observation and research strategies. To evaluate team characteristics, like the optimal timing of substitutions and tactical decisions, Hirotsu and Mike Wright [3]; [4] made use of Markov's chain theory. In addition, during the 2006 FIFA World Cup®, team tactical features were researched by Pfeiffer, Hohmann and Buehrer [5]. At present, increasingly large amounts of data in sports are collected, which creates a need for innovation in research methods. Since the use of database became an integral part in sports research, data mining methods (e.g. association rule) were first applied in table tennis [6]. In football games, association rules models have been introduced to analyze football techniques [7] and tactics [8], but most research projects focused on the development of data mining algorithm instead of its application.

In this case study, data mining methods (model of association rules) were applied to analyze the tactical behavior during all phases of ball possession for both teams (Manchester United and FC Barcelona) in the European Champions League final 2011.

2. Methods

2.1. Data collection

The match took place on May 29th, 2011, and was recorded by Video so that all game elements during the ball possession phases of the two opposing teams could be edited afterwards and used to a) analyze the game and b) compare the tactical behavior during offense.

2.2 Division of soccer field areas

As shown in Figure 1, the whole playing field is divided into 30 zones. In the first half, Manchester United attacked from right to left so their attacking area included zones #1-10. The midfield area consisted of

[+] Tianbiao Liu. Tel.: +49-921-55 3479; fax: +49-921-55 5806
E-mail address: ltbvane@yahoo.com.cn.

zones #11-20, and the backfield area of zones #21-30. Each of the three areas covered the same amount of space, 35 m out of a total of 105 m. In contrast, FC Barcelona attacked from left to right so their attacking area referred to zones #21-30, the midfield area consisted of zones #11-20, and the backfield area represented zones #1-10. In order to facilitate the recording after the teams changed ends at halftime, the numbering of the zones of the playing field remained unchanged.

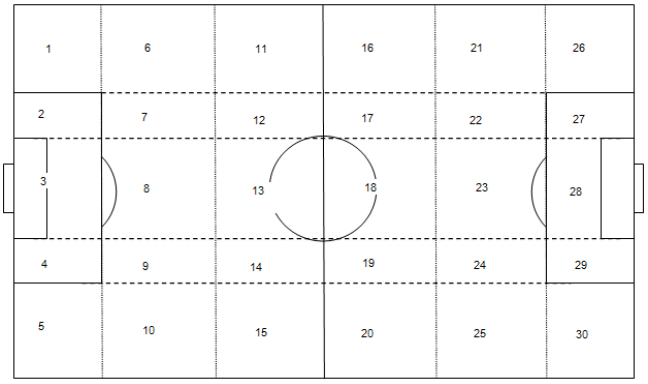

Fig. 1: Spatial division of soccer field areas

In addition to the three horizontal zones mentioned, a vertical stripe was created by zones #2-4, 7-9, 12-14, 17-19, 22-24, and 27-29. Furthermore, on both sides of this longitudinal pathway two side wings were identified by zones #1, 6, 11, 16, 21, 26, and # 5, 10, 15, 20, 25, and 30, respectively.

2.3 Data mining

2.3.1 Datamining theory

Data mining refers to extracting or "mining" knowledge from large amounts of data [9]. It is the kind of technology which extracts the non-ordinary, implicit, unknown, potentially useful information from the large-scale data [1],[11].

Data mining involves six common classes of tasks [12]: (1) anomaly detection, (2) association rule learning, (3) clustering, (4) classification, (5) regression, and (6) summarization. The association rule procedure has been used in sport events analysis to detect frequent playing patterns [13], since teams or players tend to exhibit relatively stable playing patterns. For example, teams may have players who focus on organizing the attacks, and forwards who often create one on one situation by heading into certain areas of the attacking area.

2.3.2 Association rule and Apriori algorithm

The problem of association rule mining is defined as: Let $I = \{i_1, i_2, \ldots i_m\}$ be a set of binary attributes, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I, Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction D if s% of transactions in D contains $X \cup Y$ [14].

Many efficient and scalable algorithms have been developed for frequent item set mining and the Apriori algorithm is a seminal and basic algorithm for Boolean association rules [xv]. Other improved algorithms all come from the basic Apriori. The Apriori employs an iterative approach known as a level-wise search, where k-item sets are used to explore (k+1)-item sets. First, the set of frequent 1-item sets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next L1 is used to find L2, the set of frequent 2-item sets, which is used to find L3, and so on, until no more frequent k- item sets can be found. The finding of each Lk requires one full scan

of the database.1

To improve the efficiency of the level-wise generation of frequent item sets, an important property called the a priori property is used to reduce the search space: *All nonempty subset of a frequent item set must also be frequent.*

2.3.3 Data structure

During ball possession phases the players dribble and keep passing the ball to each other, which are important actions of ball movement into the offensive area in front of the opponent`s goal. Therefore, the main interest of this study was to find out which players are important for each team's playing style, and which players serve as critical passing connections in the tactical stream of actions of the two teams.

In our research, every piece of detailed information of the game flow on the field was recorded. A team's control began when the team took possession of the ball (kick-off, free kick, throw-in, corner kick, goal kick, goalkeeper throw of the ball after the opponent's attack, possession of the ball from a tackle or one on one situation), and ended with a shot on the goal, foul, or loss of control of the ball. The data structure was as follows:

{Offensing sequence; Event in the sequence; Player number; Zone;}

The offensive sequence is the number of an offensive chain. An event in the sequence is each action number in the chain. The player number is the player who controls the ball at the moment. The zone is where the ball appears at the moment. As this structure is built, data mining can be processed to discover interesting features.

2.3.4 Application of the data mining tool

Fig. 2: Microsoft Data Mining Add-ins for Excel 2007

Table 1: part of game database using vertical format

Offensive sequence -SID	Event in the sequence - EID	Player NO. – Item set	Zone – Item set
1	1	22	6
1	2	10	11
1	3	16	12
1	4	2	14
1	5	2	20
2	1	1	3
2	2	14	5
2	3	14	10
2	4	1	3
2	5	14	10
3	1	22	11
3	2	16	11
3	3	22	11
4	1	22	6
4	2	16	6

[1] For detailed information about Apriori algorithm see Data mining: Concepts and Techniques (Han& Kamber, 2006).

A controlling chain flow (attacking flow) was developed to analyze the two team's playing patterns. In this case Microsoft Data Mining Add-ins for Excel 2007 was applied (see Fig. 2). To match the mining tool vertical data format (see Tab. 1) was applied.

In Microsoft association rule algorithm, there are three parameters reflecting the rules:

Support, Probability, and Importance

Support, which is sometimes referred to as frequency, means the number of cases that contain the targeted item or combination of items. Only items that have at least the specified amount of support can be included in the model. In contrast, the threshold for rules is expressed not as a count or percentage, but as a **probability**, sometimes referred to as confidence (see 2.3.2). Besides support and confidence (here in MS tool called probability) for each item set, the algorithm then creates scores that represent support and confidence. These scores can be used to rank and derive interesting rules from the item sets, which indicates its **importance** [15].

3. Results

3.1. Objectivity of the game observation models and model validity

In this study, all match events included in the game analysis were examined in regard to inter-rating consistency of two observers (inter-observer agreement) that was quantified by Cohen's Kappa. Manchester United first half record was selected for the examination. The Cohen's kappa values (κ) of the models were found to be: $\kappa = 0.766$ for "player's number" and $\kappa = 0.625$ for "zone". These values indicate that κ is sufficient and worthwhile to use [16], [17]. However, since the κ -values are not perfect, the difficulties of collecting match data from video tapes are underlined.

3.2. Applying Association rule to analyze both teams' control and offensive actions

After establishing the association rules model, the match data of Manchester United were mined for the first half of the game. In the data mining process, the threshold value of minimum *support* was given 5%; the threshold of minimum *probability* (confidence) was fixed at 50 %.

3.2.1 Player passing model for Manchester United (First half)

As shown in Table 2, high probabilities appeared mostly with the two center backs (#5, Ferdinand, and #15, Vidic) meaning that when one center back held the ball and then passed it to his teammate, it appeared that the ball would come back to another center back within only a few steps. The high value of importance indicates that these rules are of good quality.

Table 2: Association rule: Player number, which suggests player connections

Probability	Importance	Rule
83 %	0.63	5 = Existing, 1 = Existing -> 15 = Existing
83 %	0.63	15 = Existing, 1 = Existing -> 5 = Existing
71 %	0.30	5 = Existing, 15 = Existing -> 1 = Existing
64 %	0.63	15 = Existing -> 5 = Existing
64 %	0.63	5 = Existing -> 15 = Existing
64 %	0.29	10 = Existing -> 11 = Existing
60 %	0.24	13 = Existing -> 11 = Existing
56 %	0.46	25 = Existing -> 15 = Existing
56 %	0.46	25 = Existing -> 5 = Existing
56 %	0.19	25 = Existing -> 1 = Existing
56 %	0.19	25 = Existing -> 11 = Existing
55 %	0.20	5 = Existing -> 1 = Existing
55 %	0.20	15 = Existing -> 1 = Existing
50 %	0.41	13 = Existing -> 3 = Existing
50 %	0.30	20 = Existing -> 16 = Existing
50 %	0.14	20 = Existing -> 11 = Existing

Figure 3 shows clearly and directly Manchester United`s control of the ball in the first half of the match. From the chart one can see that the midfield players of Manchester United tend to have more vertical connections with their full backs instead of passing patterns between the midfield players themselves. That means that when one Manchester United midfield player possessed the ball and passed out, after a few steps the ball tended to go back to Manchester United defenders instead of being continuously controlled by the midfield players. One reason for that could be that the Manchester United midfield was limited by the tactical play of FC Barcelona disrupting the flow of passes which left the Manchester United midfield players with mainly one other option: vertical passes.

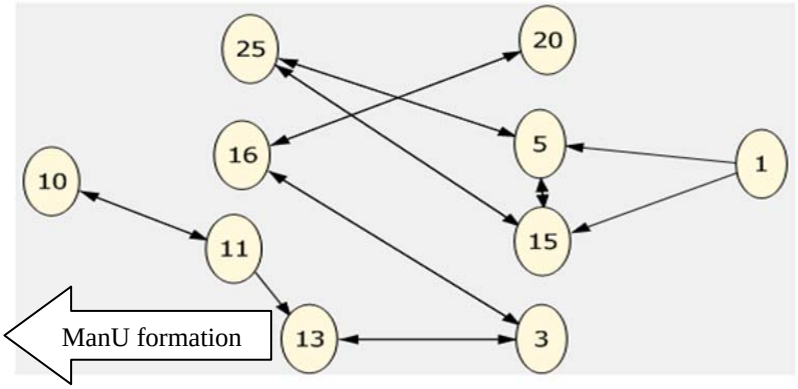

Fig. 3: Dependency network: Player passing trend of Manchester United (first half)

3.2.2 Ball movement model from area to area for Manchester United (First half)

Table 3: Association Rules: Connection between zones for Manchester United (First half)

Probability	Importance	Rule
100 %	0.77	9 = Existing -> 10 = Existing
75 %	0.67	8 = Existing -> 13 = Existing
50 %	0.35	22 = Existing -> 28 = Existing
50 %	0.52	14 = Existing -> 13 = Existing
50 %	0.35	17 = Existing -> 18 = Existing
50 %	0.60	16 = Existing -> 11 = Existing
50 %	0.60	17 = Existing -> 11 = Existing
50 %	0.38	19 = Existing -> 28 = Existing
50 %	0.38	19 = Existing -> 18 = Existing
50 %	0.52	17 = Existing -> 13 = Existing

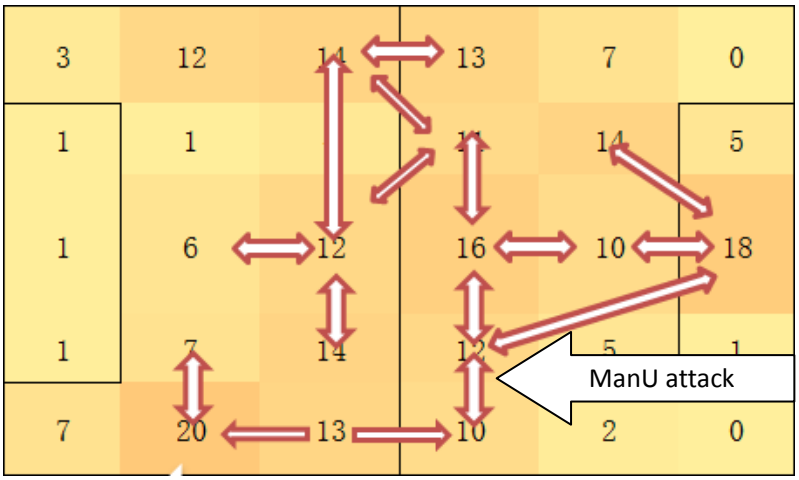

Fig. 4: Dependency Network: Zones transition for ManU (First half)

Table 3 of the first half shows that when Manchester United players controlled the ball in area #9, within the next few steps, the ball would most likely move to area #10 which turned out to be of high quality. (Importance = 0.77).

The corresponding dependency network (see Fig. 4) demonstrates more clearly the tactical differences in the Manchester United's preferred ball controlling areas. The ball could only be passed more transversely in their own half of the field compared to the number of passes into the front zones of the attacking area. This lack of vertical play was even more evident, when the ball possession took place in the front zones not closer than 35 m to Barcelona`s goal.

3.2.3 Player passing model for Manchester United (Second half)

Table 4: Association Rule: Player connections for ManU (Second half)

Probability	Importance	Rule
75 %	0.25	25 = Existing -> 10 = Existing
71 %	0.52	18 = Existing -> 15 = Existing
71 %	0.46	18 = Existing -> 25 = Existing
67 %	0.17	11 = Existing -> 10 = Existing
64 %	0.58	15 = Existing -> 1 = Existing
64 %	0.58	1 = Existing -> 15 = Existing
64 %	0.28	13 = Existing -> 3 = Existing
55 %	0.44	15 = Existing -> 13 = Existing
55 %	0.44	15 = Existing -> 5 = Existing
55 %	0.44	5 = Existing -> 15 = Existing
55 %	0.44	13 = Existing -> 15 = Existing
55 %	0.38	5 = Existing -> 25 = Existing
55 %	0.05	13 = Existing -> 10 = Existing
50 %	0.40	25 = Existing -> 5 = Existing
50 %	0.13	11 = Existing -> 3 = Existing

Table 4 shows some part of association rules of players' tactical connections within the team of Manchester United. The passes from the players #25 to #10, #18 to #15, and #18 to #25 exhibit the highest probabilities (confidence). Figure 5 presents more clearly the players connections during the second half of the match. According to this chart, Rooney played the role of the target man in Manchester United's front zone attacks.

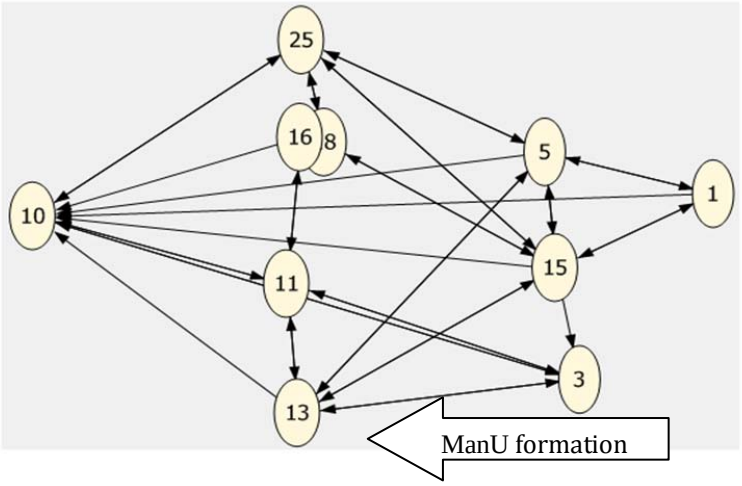

Fig. 5: Dependency Network: player passing trend of ManU (Second half)

3.2.4 Ball movement model from area to area for Manchester United (Second half)

Table 5: Association Rules: Connection between zones for Manchester United second half

Probability	Importance	Rule
100 %	0.72	11 = Existing -> 16 = Existing
75 %	0.38	17 = Existing -> 18 = Existing
71 %	0.89	8 = Existing -> 9 = Existing
71 %	0.89	9 = Existing -> 8 = Existing
63 %	0.53	17 = Existing -> 16 = Existing
60 %	0.27	16 = Existing -> 18 = Existing
60 %	0.27	15 = Existing -> 18 = Existing
50 %	0.60	16 = Existing -> 17 = Existing
50 %	1.20	16 = Existing -> 11 = Existing

Table 5 shows the data mining results of the passing between the playing field zones for Manchester United in the second half. The highest probability (confidence) occurs for passes from area #11-#16 with decent importance (0.72), while the highest importance is registered for passes from area #16-#11 with a probability of 0.5. Figure 6 presents the dependency network of the ball possession areas in the second half of the match for the ManU team. From the hardly existing connections (as illustrated in the chart), it is easy to see that Manchester United was severely limited in its attacks during the second half.

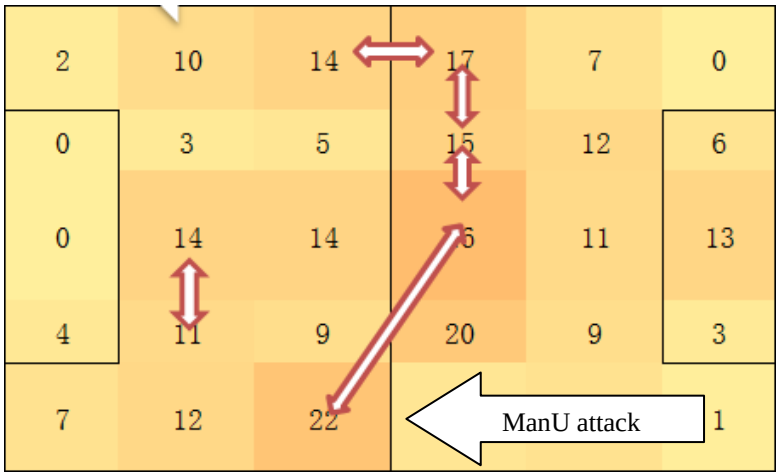

Fig. 6: Dependency network: zones transition for ManU second half

3.2.5 Player passing model for FC Barcelona (First half)

Table 6: Association Rule: Player connections for FC Barcelona (First half)

Probability	Importance	Rule
100 %	0.22	17 = Existing, 16 = Existing -> 6 = Existing
94 %	0.36	2 = Existing, 6 = Existing -> 10 = Existing
94 %	0.29	2 = Existing, 10 = Existing -> 6 = Existing
93 %	0.29	2 = Existing, 16 = Existing -> 10 = Existing
93 %	0.23	2 = Existing, 8 = Existing -> 6 = Existing
93 %	0.29	2 = Existing, 8 = Existing -> 10 = Existing
93 %	0.22	17 = Existing, 8 = Existing -> 6 = Existing
93 %	0.30	17 = Existing, 6 = Existing -> 8 = Existing
92 %	0.25	7 = Existing -> 10 = Existing
92 %	0.19	3 = Existing, 16 = Existing -> 6 = Existing
91 %	0.24	17 = Existing, 10 = Existing -> 8 = Existing

Table 6 shows parts of the association rules of the FC Barcelona players' passing trend. When #17 (Pedro) had the ball, within a few passes, #16 Busquets most probably got the ball, and then midfielder #6 (Xavi) had the highest possibility of receiving the ball.

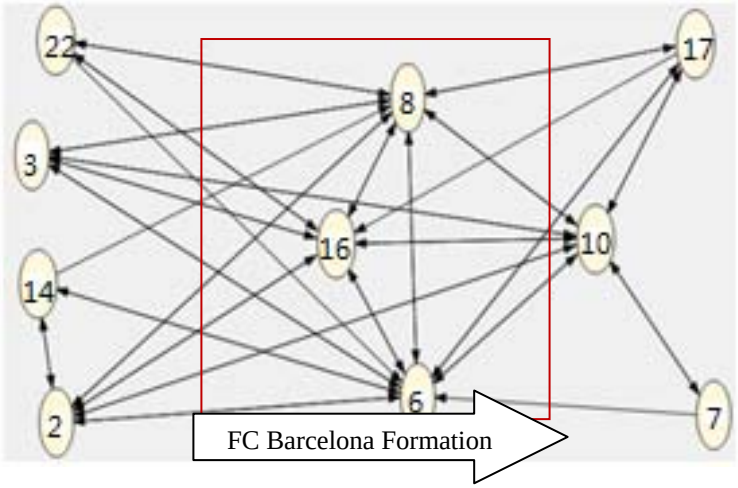

Fig. 7: Dependency network: player passing trend of FCB (First half)

Figure 7 shows FC Barcelona's connection trend. The FC Barcelona midfielders played important roles in the team's control and attack. Both offensive players (#17 and #7) on the wing sides had connections mainly with midfield players, and the fullback players were primarily connected to midfielders, too, but in no way directly to the forwards (#7 and #17) . The midfielders #6 (Xavi), #8 (Iniesta) and #10 (Messi) organized the attack in a way that is typical for FCB's playing pattern.

3.2.6 Ball moving model from area to area for FC Barcelona (First half)

Table 7: Association Rules: Connection between zones for FCB (First half)

Probability	Importance	Rule
72 %	0.51	13 = Existing -> 18 = Existing
72 %	0.51	18 = Existing -> 13 = Existing
67 %	0.35	23 = Existing -> 18 = Existing
56 %	0.41	18 = Existing -> 23 = Existing

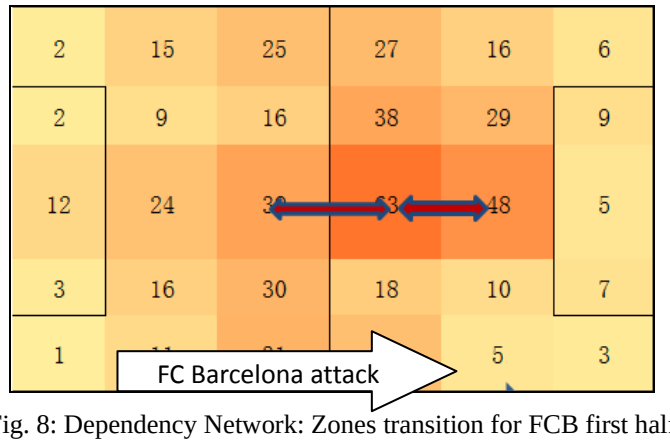

Fig. 8: Dependency Network: Zones transition for FCB first half

Table 7 shows the data mining results of the zones connections in the FC Barcelona first half match play, where the areas #13 and #18; #18 and #13; #23 and #18; #18 and #23 are characterized by high probabilities (confidence) of ball possessions. Hence the association rules for passes from area #13 to #18 and from #18 to #13 show a relative high quality. Figure 8 presents the dependency network of the different playing zones for the FC Barcelona team in the first half of the match. The chart demonstrates the FC Barcelona's controlling and passing trends that were predominantly focused on the central positions.

3.2.7 Player passing model for FC Barcelona (second half)

Table 8: Association Rule: Player connections for FCB (Second half)

Probability	Importance	Rule
100 %	0.32	22 = Existing, 6 = Existing -> 10 = Existing
100 %	0.29	8 = Existing, 2 = Existing -> 6 = Existing
94 %	0.32	8 = Existing, 10 = Existing -> 6 = Existing
91 %	0.23	22 = Existing, 10 = Existing -> 6 = Existing
91 %	0.28	8 = Existing, 2 = Existing -> 10 = Existing
89 %	0.33	8 = Existing -> 6 = Existing
88 %	0.35	8 = Existing, 6 = Existing -> 10 = Existing
88 %	0.28	2 = Existing, 10 = Existing -> 6 = Existing
85 %	0.26	22 = Existing -> 10 = Existing
85 %	0.43	10 = Existing -> 6 = Existing
84 %	0.34	8 = Existing -> 10 = Existing
79 %	0.28	2 = Existing -> 6 = Existing
79 %	0.28	2 = Existing, 6 = Existing -> 10 = Existing
79 %	0.51	6 = Existing -> 10 = Existing
77 %	0.14	22 = Existing -> 6 = Existing

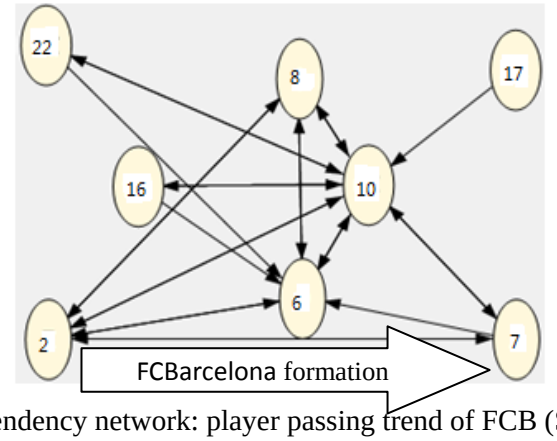

Fig. 9: Dependency network: player passing trend of FCB (Second half)

Table 8 and Figure 9 display the data mining results of the player passing model for FC Barcelona in the second half of the match. The triplets of passes from #22 to #6 to #10, and from #8 to #2 to #6 represent the main passing trends. The dependency network clarifies the tactical combinations among the players. Besides the typical playing style along the vertical midfield axis, combinations on the right wing occurred more often.

3.2.8 Ball movement model from area to area for FC Barcelona (Second half)

Table 9 and Figure 10 show the data mining results for FC Barcelona in the second half of the match. The passing trends from zone #15 to #14, #19 to #18, and from #20 to #18 occur with highest probability, indicating that these three rules are of relative high importance. In the dependency network (see Fig. 10) it can be seen that the tactical connections between FC Barcelona's midfield zones were still strong, and the right side appeared in more cases compared to the first half of the match. This finding is supported by the analysis in chapter 3.2.7, which points out that player #2 was very active on the right side in the second half of the match.

Table 9: Association Rules: Connection between zones for FCB (Second half)

Probability	Importance	Rule
86 %	0.61	15 = Existing -> 14 = Existing
85 %	0.56	19 = Existing -> 18 = Existing
83 %	0.51	20 = Existing -> 18 = Existing
77 %	0.47	19 = Existing -> 14 = Existing
72 %	0.56	14 = Existing -> 18 = Existing
72 %	0.56	18 = Existing -> 14 = Existing
71 %	0.38	12 = Existing -> 13 = Existing
71 %	0.43	15 = Existing -> 18 = Existing
67 %	0.83	14 = Existing -> 15 = Existing
67 %	0.40	18 = Existing -> 13 = Existing
63 %	0.42	13 = Existing -> 18 = Existing
61 %	0.79	18 = Existing -> 19 = Existing
56 %	0.75	18 = Existing -> 20 = Existing
56 %	0.63	14 = Existing -> 19 = Existing
56 %	0.23	14 = Existing -> 13 = Existing
56 %	0.53	18 = Existing -> 15 = Existing
53 %	0.50	13 = Existing -> 12 = Existing
53 %	0.24	13 = Existing -> 14 = Existing

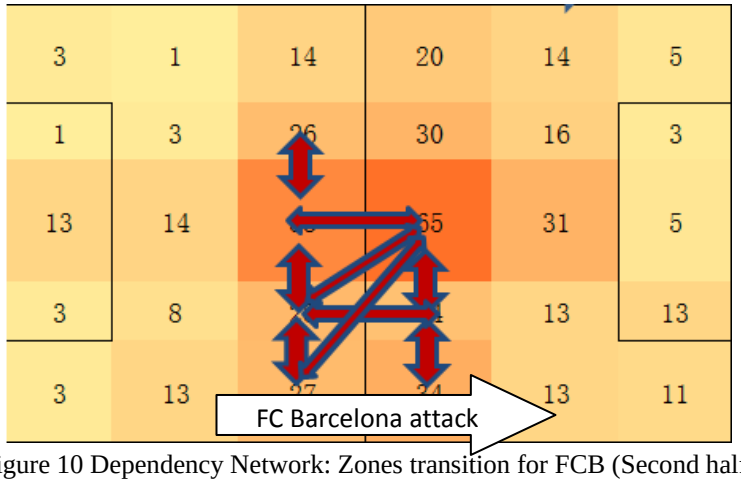

Figure 10 Dependency Network: Zones transition for FCB (Second half)

4. Discussion and conclusion

In this study, data mining theory was applied to analyze the attacking actions of the two opposing teams FC Barcelona and Manchester United in the European Champions League final 2011. By using an association rules data mining tool, both teams' basic playing patterns during their ball possession phases were clearly shown through dependency networks.

Obviously, FC Barcelona relied much more on its midfield players from the beginning of the match until the very end. Especially in the first half Manchester United's midfield was thoroughly kept in check. Although Manchester United changed its attacking strategy after the break and focused more on its forward Rooney, the English midfield vows continued as the team of FC Barcelona stayed in firm control of its midfield until the end of the match.

5. References

[1] Kuhn, Werner and Schmidt, Werner. *Analyse und Beobachtung in Training und Wettkampf: Beitraege und Analysen zum Fussballsport IV*. Sankt Augustin: Academia Verlag, 1991.

[2] Jens. Bangsbo, Tomas. Reilly, and Charles. Hughes. *Science and football*. Spon Press, 1997.

[3] N. Hirotsu, and M.Wright. An evaluation of characteristics of teams in association football by using a Markov process model. *The Statistician*. 2003, **52**: Part 4, pp. 591–602

[4] N. Hirotsu, and M. Wright. Using a Markov process model of an association football match to determine the optimal timing of substitution and tactical decisions. *Journal of the Operational Research Socitey*. 2002, **53**: 88-96.

[5] M. Pfeiffer, A.Hohmann. and M. Buehrer. Computersimulation zur Bestimmung der Leistungsrelevanz taktischer Verhaltensweisen bei der FIFA WM 2006.5. Dvs-sportspiel-Symposium Universitaet Flensburg. 2006.

[6] L. Yu, H. Zhang, and J.Hu. Computer diagnostics for the analysis of table tennis matches. *International Journal of Sports Science and Engineering*. 2008, **2**(03): 144-153.

[7] B. Wang, Z. Yin, and L. Wang. Research of Association Rules in Analyzing Technique of Football Match. 2nd *International Conference on Power Electronics and Intelligent Transportation System*. 2009.

[8] C. Pan. Appliance of Apriori Algorithm on technical-tactics analysis of football. *Computer Knowledge and Technology*. 2010, **31**(6): 8835-8837.

[9] J. Han. and M. Kamber. *Data Mining: Concepts and Techniques* (Second Edition), Elsevier Inc, 2006.

[10] J, Brachman, and T,Anand. The Process of Knowledge Discovery in Databases. *A Human centered Approach*. 1996, pp. 56-60.

[11] P. Smyth. From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*.1991, pp. 56-60.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery in Databases. Retrieved 2008-12-17.

[13] I. Bhandari, E. Colet, and J. Parker. Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery*. 1997, **1**: 121–125,

[14] Agrawal and Srikant. *Fast Algorithms for Mining Association Rules*. 1994.

[15] J. Han, and M. Kamber. *Data Mining: Concepts and Techniques* (Second Edition). Elsevier Inc, 2006.

[16] MSDN, Microsoft Association Algorithm Technical Reference. http://msdn.microsoft.com/en-us/library/cc280428.aspx

[17] W. Greve, and D.Wentura. Wissenschaftliche Beobachtung: Eine Einführung, PVU/Beltz, Weinheim 1997.

[18] J. Landis, and G. Koch. The measurement of observer agreement for categorical data. In: *Biometrics*. 1977, **33**: 159–174.