The Application of Data Mining in Sports Events

Jiao Yu-ping[1,a], Li Xia[2,b]

[1]Physical Education Department，Guangdong University of Foreign Studies, China

[2]Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, China

[a]michael0924@qq.com,[b]200211025@oamail.gdufs.edu.cn

Key words: Data mining; The Asian Games; Volleyball; Correlation analysis

Abstract.

In recent years, data mining has played a more and more important role in all fields and had an increasingly greater influence. By using association analysis of data mining and classification algorithm, this paper analyzes the result data of volleyball games, which were held in a university's volleyball venue during 2010 Guangzhou Asian Games. It studies the correlation between different competing countries, competition time and audience number and make correlation analysis between the competition results and athletes' physical qualities. These results may be significant for scientific game guidance and athlete selection.

1. Introduction

With the development of China's competitive sports and the constant improvement of athletes' competitive levels, a higher request was also put forward towards sports training and competition efficiency. Today, in the age of Internet and technology, both the training of competitive sports and the selection of athletes need a more scientific prediction, management and program. However, the data management of our country's competitive sports is largely at a state of disorder at present: modern information technology and various data of country's competitive sports development, which was accumulated for years, have not been fully used; regularities and modes other than objective experience have not been mined efficiently neither. In these 10 years, the skill of data mining and analysis becomes increasingly mature. If it is applied to the training and the competing process of athletic competition, they will be more scientific and standardized.

Data mining is defined as mining the hidden, unknown but potentially useful information from plenty of actually applied data, which is massive, incomplete, noisy, ambiguous and random. These hidden, unknown, but potentially useful information can be presented in various forms such as concept, rule, mode and law. Simply put, data mining is a deep-level data analysis approach. That is: data mining is a kind of information technology which is not limited to search and access, but find potential links between different data. This paper analyses the result data of volleyball games, which was held in a volleyball venue of a university during 2010 Guangzhou Asian Games, using Apriori algorithm and ID3 algorithm. It studies the correlation between different competing countries, competition time and audience number and the correlation between competition results and athletes' physical qualities respectively. These results may be significant for scientific game guidance and athlete selection.

2. The application of data mining in the data analysis of Asian Games' volleyball matches

2.1 Experimental data

All data for this research comes from the information department of main volleyball venue of women volleyball games at 2010 Asian Games (Guangwai Gymnasium). All data is realistic and believable including detail information such as weight and height of athletes from 11 competing nations and statistics of 35 matches.

2.1.2 Data pre-processing

The data of this article comes from the volleyball’s group game record in 2010 Asian Games, Guangwai Gymnasium. The record contains 5 data field: Num, Country1, Country2, Time and Class. 35 data items was included in total and 11 competing country was included such as KAZ, TPE and CHN. In order to facilitate mining, we pre-process the collected data.

3 Analysis of experimental results

3.1 Association analysis of competing nations, audience number and game time

Mining pre-processed arff documents with association analysis method using WEKA software. Results are as follow.

```
Apriori
= = = = = = =
Minimum support : 0.1 (4 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 15
Size of set of large itemsets L(2): 7
Best rules found:
1. Country1=MGL 4 = => time=1    4 conf:(1)
2. Country1=CHN 4 = => class=3    4 conf:(1)
3. Class=3 5= =>country1=CHN 4    conf:(0.8)
4. Countey2=MDV 5 = =>time=1    4 conf:(0.8)
```

Analysis of experimental results：

On the basis of the 4 association rules from the experiment results, we can reach the following conclusion.

(1) When CHN team takes part in the match, audience will always reach a number of 1500 or above. As a result, organizing committee should arrange more hands in volunteering, security and logistical support to avoid accident like stampede, which happens because of over-crowding.

(2) When MGL (Mongolia) or MDV (Maldives) team takes part in a match, competing time of the match will always limited to 40-60 minutes because of the weak strength of those teams. So under this circumstance, organizing committee should arrange corresponding staff to hold their position as quick as possible to fit in the short competing time and make good preparation for the next match.

3.2 Data mining analysis of single technique’s rank and athlete’s physical qualities based on correlation analysis

3.2.1 Experimental objectives

By concluding from the scores of 35 women's volleyball matches, in order to find the rules which can be used as an index for athlete’s selection，we hope that a relation between physical quality and scores could be confirmed through the correlation analysis of some related data of the physical quality about the top 48 athletes in single technic.

3.2.2 Introduction of data set and data pre-process

Description of this experiment’s raw data set and data pre-process in this paper are shown as table1, 2 and 3. This research spent a large amount of time on the pre-process of the data.

Table1. Data3—Athlete’s technical ranking of Asian group games.txt

Athlete’s technical ranking of Asian group games

Data sources	Technical ranking of volleyball athletes on Guangwai volleyball venue in 2010 Asian Games			
Data pre-process	We double-checked, cleared and corrected null value and default on the basis of original hard copy file data.			
Number of items	48	Only those who ranks inside top 48 with total technical score are recorded		
Fields included	7			
Abbreviation	Name of data fields	Type	Note	Range
rk	rank	figure		1~48
no	number	category		1~13
noc	nation	category	KAZ=Kazakhstan, TPE=Taipei (China), PRK=People's Republic of Korea, THA=Thailand, TJK=Tajikistan, MGL=Mongolia, MDV=Maldives, KOR=Korea, JPN=Japan, IND=India, CHN=China	
spike	spike	figure	times of athletes' spike	16~120
block	block	figure	times of athletes' block	0~21
serve	serve	figure	times of athletes' serve	0~19
total	total	figure	total scores of the three skills	27~139

Table2 Data4--- Data of athlete's physical quality.txt

Data of athlete's physical quality				
Data sources	Athlete's physical quality data in the group games on Guangwai volleyball venue in 2010 Asian Games			
Data pre-process	We double-checked, cleared and corrected null value and default on the basis of original hard copy file data. For some missing official data, we use "?" to represent.			
Number of items	130	The data are from the athletes who come from 11 participating countries respectively in this Asian Games, women's volleyball games.		
Fields included	7			
Abbreviation	Name of data fields	Type	Note	Range
noc	nation	category	KAZ=Kazakhstan, TPE=Taipei (China), PRK=People's Republic of Korea, THA=Thailand, TJK=Tajikistan, MGL=Mongolia, MDV=Maldives, KOR=Korea, JPN=Japan, IND=India, CHN=China	
no	number	category		
height	height(cm)	figure	athlete's height (cm)	167~192
weight	weight(kg)	figure	athlete's weight	61~90
blockmax	the highest point of spike (cm)	figure	the max height of spike that the athlete can reach	260~320

servemax	the highest point of block (cm)	figure	the max height of block that the athlete can reach	260~320
b**irth**	year of birth	figure	athlete’s year of birth	1978~1994

Table3 Data5---- Comparison of technical ranking and athlete’s physical index.arff

Comparison of technical ranking and athlete’s physical index				
Data sources	Two data sources that table1 and table2 corresponds with respectively			
Data pre-process	Data from table1 and table2 was collected and data transformation was done to adapt the data set to WEKA format.			
Number of items	48	The data are from the athletes who come from 11 participating countries respectively in this Asian Games, women’s volleyball games.		
Fields included	12			
Abbreviation	Data field name	Type	Note	Range
rk	rank	figure		1~48
no	number	category		1~13
noc	nation	category	KAZ=Kazakhstan, TPE=Taipei (China), PRK=People’s Republic of Korea, THA=Thailand, TJK=Tajikistan, MGL=Mongolia, MDV=Maldives, KOR=Korea, JPN=Japan, IND=India, CHN=China	
spike	spike	figure	times of athletes’ spike	16~120
block	block	figure	times of athletes’ block	0~21
serve	serve	figure	times of athletes’ serve	0~19
total	total score	figure	total scores of the three skills	27~139
height	height(cm)	figure	athlete’s height (cm)	167~192
weight	weight(kg)	figure	athlete’s weight	61~90
spikemax	the highest point of spike (cm)	figure	the max height of spike that the athlete can reach	260~320
blockmax	the highest point of block (cm)	figure	the max height of block that the athlete can reach	260~320
birth	year of birth	figure	athlete’s year of birth	1978~1994

After organizing needed data sheets: table1, 2 and 3, in order to mine the results of this experiment (relationship between athlete’s single skill ranking and physical quality), athlete’s height, weight, spikemax, blockmax and birth were extracted from table2 (data of athlete’s physical quality). Along with table2’s data, correlated single skill rankings in table1 were extracted as well to form the 48 samples. Because the data are used to make correlation analysis with Apriori algorithm, 6 attributes of the data set that we prepared are figures; we need to divide them into different sections to adapt this algorithm. We divided the data into 4-6 sections by rules respectively. Data51.arff are produced as belows.

Table4. Data51---- Comparison of technical ranking and athlete’s physical index.arff

Comparison of technical ranking and athlete’s physical index

Data sources	Data from table1 and table2			
Data pre-process	For the fact that the data set itself has a mass of number attribute, we pre-treat several attributes into sections in order to make correlation analysis with Apriori algorithm.			
Number of items	48	The data are from the athletes who come from 11 participating countries respectively in this Asian Games, women’s volleyball games.		
Fields included	6			
Abbreviation	Name of data fields	Type	Note	Range
height	athlete’s height (cm)	category	athlete’s height (cm)	1,2,3,4,5,6
weight	athlete’s weight	category	athlete’s weight	1,2,3,4,5,6
spikemax	the highest point of spike (cm)	category	the max height of spike that the athlete can reach	1,2,3,4,5,6
blockmax	the highest point of block (cm)	category	the max height of block that the athlete can reach	1,2,3,4,5,6
birth	year of birth	category	athlete’s year of birth	1,2,3,4
class	ranking class	category	athlete’s ranking section	1,2,3,4,5,6

The expression of data section

height	weight	spike	block	birth	total	
166-170	61-65	261-270	261-270	76-80	20-39	1
171-175	66-70	271-280	271-280	81-85	40-59	2
176-180	71-75	281-290	281-290	86-90	60-79	3
181-185	76-80	291-310	291-300	91-95	80-99	4
186-190	81-85	301-310	301-310		100-119	5
191-195	86-90	311-320	311-320		120-139	6

3.2.3 Mining data by WEKA and Apriori algorithm

WEKA and Apriori algorithm were used to mine the data in table 3 and 4 In the process of mining, the confidence coefficients of these two algorithms were set as 0.9 and 0.7 respectively (due to space constraints, we haven’t provide the experimental conclusion but only analysis.) In the end, experimental conclusion was drawn.

(1) There is no strong relationship between physical quality and competition results. Athlete around 25 years old not only owns an excellent physical fitness but also have abundant competition experience, which is helpful to adjust their psyche in the competition.

(2) Although there is no strong relationship between physical quality and competition results, we can see that height enjoy an advantage in volleyball match. Athletes with appropriate weight and height serve more powerful, and their blocks will be relatively a lot easier. It is a must for them to achieve good results.

(3) Therefore, when selecting athletes, it’s more favorable for the improvement of competition results to choose athletes with about 185cm in height and 70kg in weight.

(4) The disadvantage of this experiment is that mining result may have errors due to the less data set.

3.3 Data mining of differences between Athletes' physical conditions of each competing nations on the basis of correlation analysis

3.3.1 Experimental objectives

To make some objective recommendation to the training and selecting of athletes through analyzing data such as height, weight, spike and block of athletes from 11 competing countries.Data pre-process:Put hard-copy data into computer; delete some items without athlete’ s personal information and transfer some field types (transformation towards the inter-partition). One of the original 11 countries was also deleted because it never took part in international games before and its athletes’ information was absent. Therefore, 10 countries’ athlete information was left.

3.3.2 Data mining of correlation analysis

Mining the data that shows the differences in physical quality of different nations' athlete using Apriori algorithm, we can draw the following conclusion:

(1) The heights of spiking and blocking are strongly correlated, and it can be confirmed that the height of the players is related to their level of blocking by changing the parameters.

(2) The heights of spiking and blocking are also related to the weight of the players. Generally, lightweight players (≤64 kg) can reach the heights of no more than 289cm when spiking and no more than 279cm when blocking. As a result, players who are comparatively higher and lighter should be a better choice for training and selecting.

Summary

Data mining is a new information technology with strong vitality and actual effect, which can be widely applied to all fields. Among them, the field of sports has accumulated massive data. So there is no doubt that sports statistic plays an irreplaceable role in scientific studies of sports. The data mining method that was put forward in this paper can not only be applied to technical-tactics analysis and staffing arrangement of volleyball games, but also to the analysis and application of other events. Data mining will exert a greater influence on all fields in the future. Also, it will have a broad application prospect on the field of sports.

References

[1] Pawan Lingras,Unsupervised Rough Set Classification Using Gas, Journal of Inteligent Information Systems, 2001 ,16: 215~228 .

[2]L.P.Khoo,S.B.Tor and LY.Zhai,A Rough-Set-Based Approach for Classification and Rule Induction,International journal of Advanced Manufacturing Technology, 1999, 15 :438-444

[3]Wu X, Zhu X,Wu G Q, ea al.Date mining with big date[J].Knowledge and Date Engineering,IEEE Transactions on ,2014,26(1):97-107.

[4] Bryant R, Katz R H,Lzaowska E D. Big-Data Computing: creating Revolutionary Breakthroughs in Commerce, Scieence and Society[J].2008.

[5]McKenna A, Hanna M, Banks E, et al.The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generationg DNA sequencing date [J].Genome research, 2010,20(9):1297-1303.

[6]Aleksander Ohrn.ROSETTA Technical Reference Manua.Knowledge Systems Group. Department of Comouter and Information Science, NTNU.Trondheim,Norway.1999(11).