# Analysis of NBA Player Positions Using ISODATA: A Case Study

CHE-CHERN LIN[1], JHIH-YANG CHEN[2], YI-TING CHOU[3], YUN-HAO HUANG[4],
YUN-YI TONG[5], HOWARD LO[6]
Department of Industrial Technology Education
National Kaohsiung Normal University
TAIWAN

*Abstract:* - In this paper, we apply the ISODATA to cluster NBA player positions according to their game statistics in regular season 2005-2006. In the experiment, there were totally 214 players used to be examples. Three phases were processed during the experiment based on three different desired clusters (positions). In the first phase, three positions were classified. In the second phase, five positions were classified. In the third phase, seven positions were classified. We used confusion matrices to calculate the clustering accuracies. The experimental results show that less numbers of the desired clusters get better clustering results.

*Key-Words:* - Classification; ISODATA; Basketball; Player position; NBA.

## 1 Introduction

Two major data clustering approaches, the K-mean algorithm and the Iterative Self-Organizing Data Analysis Technique (ISODATA), are often used to cluster samples into different groups based on statistical parameters. The K-mean algorithm assumes the number of clustering centers is given and uses an iterative method to calculate cluster centers. The ISODATA has more flexibility to determine clustering centers where means and standard deviations are utilized to cluster samples. Mainly, the ISODATA is an iterative method and uses two phases to determine cluster centers: a merging phase and a splitting phase. In the merging phase, two or more clusters are merged if the distances among these clusters are less than a threshold value. In the splitting phase, a cluster is divided into more clusters if its standard deviation is greater than a threshold value. In brief, the procedure of the ISODATA is described as follows: determine clustering parameters; assign the sample to current cluster centers; discard the clusters if their sample sizes are less than the desired value; calculate the average distance for each cluster; calculate the overall average distance for the entire sample space; find the maximum component among the standard deviations; split a cluster if necessary; merge clusters if necessary. The computational details can be found in [1].

Recently, the ISODATA is widely applied to cluster numerical data in many areas. It has been applied to increase the sensitive of Magnetic Resonance Imaging (MRI) parameters to detect the damage of tissue in ischemic stroke [2]. The ISODATA has also been utilized to reduce the computational complexity for a Gaussian-Mixed-Model-based speaker recognition system [3]. An ISODATA-based color clustering method has been proposed to extract caption segmentation from videos using topographical features where the ISODATA served to cluster similar colors. A fast implementation of the ISODATA has been presented to reduce computational time by sorting the points in kd-tree and estimation of dispersion of each cluster [4].

In this paper, we apply the ISODATA to cluster the NBA (National Basketball Association in the United States of America) players' positions based on the plays' statistics of the season of 2005-2006. The players' data are obtained form the official website of NBA. Three phases were processed during the experiment based on three different desired clusters (positions). In the first phase, three positions were classified. In the second phase, five positions were classified. In the third phase, seven positions were classified.

## 2 Background

In this section, we introduce the background of basketball. We demonstrate an illustrative figure of a basketball court and introduce basketball players' positions. We also explain the player positions shown on the official website of NBA [5].

An illustrative figure for a basketball court is displayed in Figure 1 [6,7,8,9]. When playing a game, each team has five players on the court. The five players are regularly numbered from 1 to 5 according to the locations where they play. The five players are described as follows [6,7,8,9]:

1. Player 1 (Point guard, PG): This player plays a very important role to handle and coordinate the offense. He should know everything about the game plan and make decisions to deliver the ball to an appropriate player using his excellent passing and dribbling skills.
2. Player 2 (Shooting guard): This player is primarily designed to get points during a game using his good shooting skills. He may also pass the ball to the post players (power forward and center) and therefore usually plays the wind area.
3. Player 3 (Small forward, SF): This player plays a mixing role consisting of guards and post players. He is also one of the important players to run the defense with his excellent defense skills.
4. Player 4 (Power forward, PF): This player is primarily designed to get the rebounds during a game using his strong body to physically contact the opponent players.
5. Player 5 (Center, C): This player is probably the tallest person on the team using his good catching skills and helps his team hrun the offense. On defense, He is also the last player to prevent from shooting made by the opponent.

According to the official website of the NBA, players are basically clustered into three fundamental positions: center, forward, and guard. Since some players play two positions, the NBA then uses seven positions to categorize the players. The seven positions are explained as follows:

1. C: playing a center position.
2. C-F: playing both center and forward positions but mostly playing a center position.
3. F: playing a forward position.
4. F-C: playing both forward and center positions but mostly playing a center position.
5. F-G: playing both forward and guard positions but mostly playing a forward position.
6. G: playing a guard position.
7. G-F: playing both guard and forward positions but mostly playing a guard position.

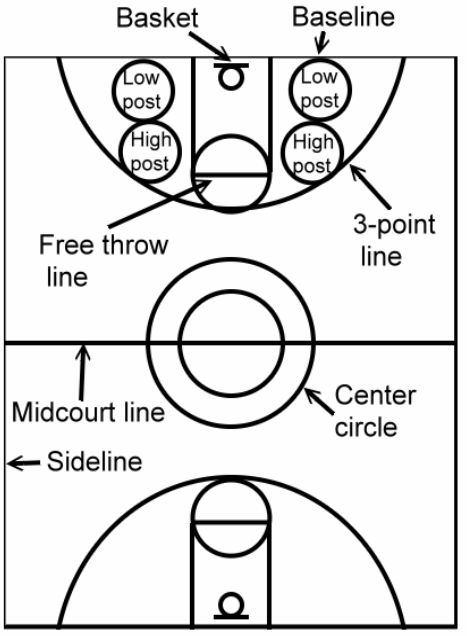


Fig. 1: An illustrative figure of basketball court (taken form [6,7,8,9])

# 3 Experimental Setups

## 3.1 Data acquirement

In this experiment, the data are obtained from the official website of NBA [5]. We selected the players' statistics of regular season 2005-2006. All of the thirty teams are considered for the experiment. There are currently 410 players in the NBA. We selected those players who played at least 30 games in the season and played at least 20 minutes (in average) per game as our experimental samples. Of the 410 players, 214 players are qualified to be the samples. Seven attributes are selected for this experiment including field goal percentage, three-point percentage, foul shot percentage, rebound, assistance, steal, and block.

## 3.2 Data normalization

The data were normalized using the min-max value normalization from a range of 10 to 255 according to the following equation [10]

$$NormValue = \frac{origValue - oldMin}{oldMax - oldMin}(newMax - newMin) + newMin \quad (1)$$

where *normValue* is the normalized value; *origValue* is the original; *oldMax* and *oldMin* are the maximum and minimum of the original data set; *newMax* (=255) and *newMin* (=10) are the maximum and minimum of the normalized data set.

### 3.3 Procedure

There are three phases in the experiment based on three different desired clusters (positions). In the first phase, three desired clusters were assigned according to three different positions including a center, forward, and guard. The data in the three desired positions are obtained from the statistical data of the seven positions demonstrated on the NBA's website [5], as mentioned early, by the following assignments:

Desired position C (center) is obtained from NBA positions of C and C-F; desired position F (forward) is obtained from NBA positions of F, F-C and F-G; desired position G (guard) is obtained from NBA positions of G and G-F.

In the second phase, the desired positions are the five regular positions: PG (Player 1), SG (Player 2), SF (Player 3), SF (Player 4), and PF (Player 5).

In the third phase, the desired positions are exactly the same as the NBA's seven positions.

## 4 Experiment Results

We used confusion matrices to evaluate the experiment results. The diagonal items in a confusion matrix represent the numbers of samples which are correctly clustered. However, the off-diagonal ones represent the numbers of samples which are mis-clustered. During experiment, the computed cluster numbers varied depending on the parameters. If the number of the computed cluster centers is larger than the number of desired cluster centers we then assigned the computed clusters to appropriate the desired clusters (positions) for calculating accuracy in a confusion matrix. For examples, if the number of desired positions is three (positions C, F, G) and the number of computed clusters is five, we need to assign the five computed clusters to the three desired positions to establish a confusion matrix to calculate the clustering accuracy. Tables 1 to 12 show the experimental results. Figure 2 shows the experimental summary.

**Table 1: The confusion matrix of 3 desired clusters *v.s.* 3 computed clusters.**

| Desired | Computed results | | | |
|---|---|---|---|---|
| clusters | C | F | G | total |
| C | 27 | 18 | 0 | 45 |
| F | 2 | 47 | 3 | 52 |
| G | 1 | 25 | 91 | 117 |
| Total | 30 | 90 | 94 | 214 |

1. Clustering accuracy = 77.10%.

**Table 2: The confusion matrix of 3 desired clusters *v.s.* 5 computed clusters.**

| Desired | Computed results | | | |
|---|---|---|---|---|
| clusters | C | F | G | total |
| C | 25 | 18 | 0 | 43 |
| F | 4 | 52 | 11 | 67 |
| G | 1 | 20 | 83 | 104 |
| Total | 30 | 90 | 94 | 214 |

1. Clustering accuracy = 74.77%.
2. The 5 computed clusters are appropriately assigned to C, F, and G positions.

**Table 3: The confusion matrix of 3 desired clusters *v.s.* 6 computed clusters.**

| Desired | Computed results | | | |
|---|---|---|---|---|
| clusters | C | F | G | total |
| C | 26 | 18 | 1 | 45 |
| F | 3 | 45 | 3 | 51 |
| G | 1 | 27 | 90 | 118 |
| Total | 30 | 90 | 94 | 214 |

1. Clustering accuracy = 75.23%.
2. The 6 computed clusters are appropriately assigned to C, F, and G positions.

**Table 4: The confusion matrix of 3 desired clusters *v.s.* 7 computed clusters.**

| Desired | Computed results | | | |
|---|---|---|---|---|
| clusters | C | F | G | total |
| C | 25 | 18 | 0 | 43 |
| F | 4 | 53 | 7 | 64 |
| G | 1 | 19 | 87 | 107 |
| Total | 30 | 90 | 94 | 214 |

1. Clustering accuracy = 77.10%.
2. The 7 computed clusters are appropriately assigned to C, F, and G positions.

**Table 5: The confusion matrix of 3 desired clusters v.s. 10 computed clusters.**

| Desired | Computed clusters | | | |
|---|---|---|---|---|
| clusters | C | F | G | total |
| C | 24 | 16 | 0 | 40 |
| F | 5 | 59 | 6 | 70 |
| G | 1 | 15 | 88 | 104 |
| Total | 30 | 90 | 94 | 214 |

1. Clustering accuracy = 79.91%.
2. The 10 computed clusters are appropriately assigned to C, F, and G positions.

**Table 6: The confusion matrix of 3 desired clusters v.s. 14 computed clusters.**

| Desired clusters | Computed clusters C | F | G | total |
|---|---|---|---|---|
| C | 25 | 18 | 0 | 43 |
| F | 4 | 53 | 7 | 64 |
| G | 1 | 19 | 87 | 107 |
| Total | 30 | 90 | 94 | 214 |

1. Clustering accuracy = 77.10%.
2. The 14 computed clusters are appropriately assigned to C, F, and G positions.

**Table 7: The confusion matrix of 5 desired clusters *v.s.* 5 computed clusters.**

| Desired clusters | Computed results C | PF | SF | SG | PG | total |
|---|---|---|---|---|---|---|
| C | 25 | 16 | 2 | 0 | 0 | 43 |
| PF | 2 | 18 | 2 | 1 | 0 | 23 |
| SF | 2 | 15 | 17 | 7 | 3 | 44 |
| SG | 1 | 3 | 12 | 27 | 23 | 66 |
| PG | 0 | 0 | 6 | 15 | 17 | 38 |
| Total | 30 | 52 | 39 | 50 | 43 | 214 |

1. Clustering accuracy = 48.60%.

**Table 8: The confusion matrix of 5 desired clusters *v.s.* 7 computed clusters.**

| Desired clusters | Computed results C | PF | SF | SG | PG | total |
|---|---|---|---|---|---|---|
| C | 25 | 16 | 2 | 0 | 0 | 43 |
| PF | 2 | 18 | 2 | 0 | 0 | 22 |
| SF | 2 | 15 | 18 | 6 | 1 | 42 |
| SG | 1 | 3 | 11 | 28 | 22 | 65 |
| PG | 0 | 0 | 6 | 16 | 20 | 42 |
| Total | 30 | 52 | 39 | 50 | 43 | 214 |

1. Clustering accuracy = 50.93%
2. The 7 computed clusters are appropriately assigned to C, PF, SF, SG and PG positions.

**Table 9: The confusion matrix of 5 desired clusters *v.s.* 10 computed clusters.**

| Desired clusters | Computed results C | PF | SF | SG | PG | total |
|---|---|---|---|---|---|---|
| C | 24 | 16 | 0 | 0 | 0 | 40 |
| PF | 4 | 21 | 7 | 1 | 0 | 33 |
| SF | 1 | 13 | 19 | 3 | 1 | 37 |
| SG | 1 | 2 | 13 | 39 | 27 | 82 |
| PG | 0 | 0 | 0 | 7 | 15 | 22 |
| Total | 30 | 52 | 39 | 50 | 43 | 214 |

1. Clustering accuracy = 55.14%.
2. The 10 computed clusters are appropriately assigned to C, PF, SF, SG and PG positions.

**Table 10: The confusion matrix of 5 desired clusters v.s. 14 computed clusters.**

| Desired clusters | Computed clusters C | PF | SF | SG | PG | total |
|---|---|---|---|---|---|---|
| C | 24 | 16 | 0 | 0 | 0 | 40 |
| PF | 4 | 17 | 5 | 0 | 0 | 26 |
| SF | 1 | 17 | 25 | 14 | 4 | 61 |
| SG | 1 | 2 | 9 | 32 | 19 | 63 |
| PG | 0 | 0 | 0 | 4 | 20 | 24 |
| Total | 30 | 52 | 39 | 50 | 43 | 214 |

1. Clustering accuracy = 55.14%.
2. The 14 computed clusters are appropriately assigned to C, PF, SF, SG and PG positions.

**Table 11: The confusion matrix of 7 desired clusters *v.s.* 7 computed clusters.**

| Desired clusters | Computed results C | CF | F | FC | FG | G | GF | total |
|---|---|---|---|---|---|---|---|---|
| C | 14 | 1 | 0 | 1 | 0 | 0 | 0 | 16 |
| CF | 6 | 5 | 11 | 3 | 1 | 0 | 1 | 27 |
| F | 2 | 0 | 24 | 2 | 6 | 7 | 1 | 42 |
| FC | 1 | 1 | 13 | 1 | 0 | 0 | 0 | 16 |
| FG | 0 | 0 | 3 | 0 | 2 | 36 | 1 | 42 |
| G | 1 | 0 | 9 | 0 | 5 | 45 | 5 | 65 |
| GF | 0 | 1 | 5 | 0 | 0 | 0 | 0 | 6 |
| Total | 24 | 8 | 65 | 7 | 14 | 88 | 8 | 214 |

1. Clustering accuracy = 42.52%

**Table 12: The confusion matrix of 7 desired clusters *v.s.* 14 computed clusters.**

| Desired clusters | Computed results | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | C | CF | F | FC | FG | G | GF | total |
| C | 11 | 3 | 3 | 2 | 0 | 0 | 0 | 19 |
| CF | 9 | 5 | 14 | 3 | 0 | 0 | 1 | 32 |
| F | 2 | 0 | 12 | 0 | 1 | 0 | 0 | 15 |
| FC | 0 | 0 | 11 | 2 | 2 | 2 | 2 | 19 |
| FG | 1 | 0 | 17 | 0 | 8 | 15 | 1 | 42 |
| G | 0 | 0 | 3 | 0 | 2 | 49 | 1 | 55 |
| GF | 1 | 0 | 5 | 0 | 1 | 22 | 3 | 32 |
| Total | 24 | 8 | 65 | 7 | 14 | 88 | 8 | 214 |

1. Clustering accuracy = 42.06%
2. The 14 computed clusters are appropriately assigned to C, CF, F, FC, FG, G, and GF positions.

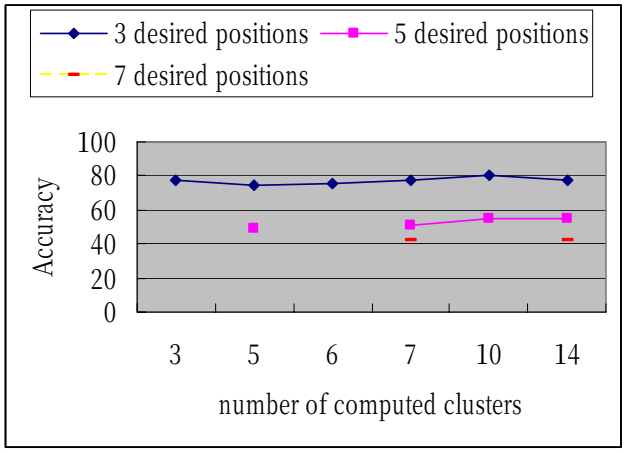


Fig. 2: The experimental summary

## 5 Conclusions

We applied the ISODATA to cluster NBA player positions based on their game statistics in regular season 2005-2006. In the experiment, there were 214 players used to be examples and three phases were processed based on different desired positions. We divided the examples into several clusters and used confusion matrices to calculate the clustering accuracies. The experimental results show that less numbers of the desired clusters get better clustering results. This ISODATA-based clustering system could be used as a support system to help basketball coaches to recruit new players with statistical supports. Two common statistical clustering algorithms are often used to get clustering centers: the ISODATA algorithm and the K-mean algorithm. It might be interesting issue to compare the clustering results of both methods.

To use more attributes and to apply more complicated methods (i.e. fuzzy clustering techniques) to correctly cluster the overlapped samples might be the directions for the future studies. Further analyses using artificial intelligence techniques such as data mining and neural networks are also good research topics for future works.

*References*

[1] J. T. Tou and R.C. Gonzalez, *Pattern Recognition Principle*, Reading, MA: Addison-Wesley Publishing Company Inc., 1981.

[2] Guangliang Dinga, Quan Jianga, Li Zhanga, Zhenggang Zhanga, Robert A. Knighta, Hamid Soltanian-Zadehb, Mei Luc, James R. Ewinga, Qingjiang Lia,Polly A. Whittona, Michael Choppa, "Multiparametric ISODATA analysis of embolic stroke andrt-PA intervention in rat," JNS 6 May 2004.

[3] Bing Sun, Wenju Liu, Qiuhai Zhong, "Hierarchical Speaker Identification Using Speaker Clustering," NLPR.

[4] Nargess Memarsadeghi, David M. Mount, Nathan S. Netanyahu, "A Fast Implementation of the Isodata Clustering Algorithm," IJCGA, December 31, 2005.

[5] NBA Official Website: http://www.nba.com.

[6] Hai Wissel, *Basketball: Step to Success*, Human Kinetics Publishers, Inc., 1994, pp. 1-3.

[7] Sandy L. Simpson, *Coaching Girls' Basketball*, Random House, Inc., 2001, pp. 41-56.

[8] John P. McCartby, Jr., *Coaching Youth Basketball 2nd Edition*, John P. McCartby, Jr, 1996, pp. 95-110.

[9] C.C. Lin, V Chen, C.C. Yu, Y.C. Lin, "A Schema to Determine Basketabll Defense Strategies Using a Fuzzy Expert Systems," The proceeding of the 7th WSEAS international conference on Fuzzy Systems, Cavtat, Croatia, June 12-14, 2006.

[10] Richard J. Roiger, Michael W. Geatz, *Data Mining a tutorial-based primer*, Addison-Wesley, 2003, pp.155-156.

[11] Josef Cihlar, Rasim Latifovic, Jean Beaubien, "A Comparison Of Clustering Strategies For Unsupervised Classification," Canadian Journal of Remote Sensing, October 22, 1999.