Data Mining and Basketball Games

Processing Game Data for the Machine Learning Process

Joel Poualeu
Department of Computer and Information Sciences
The College of St. Scholastica
JPouale2@css.edu

Advised by:

Dr. Yu-Han Chang
Computer Science Department
University of Southern California
Information Sciences Institute
ychang@isi.edu

Dr. Rajiv Maheswaran
Computer Science Department
University of Southern California
Information Sciences Institute
maheswar@usc.edu

Abstract

The various events occurring during a basketball game such as shots made and shots missed, and the factors affecting these events such as the time period when they occurred or the players on the court when the event takes place, can build up more or less regular patterns that can eventually be defined by player, team and/or year. These patterns can exist in terms of coördinate values on a typical basketball court, be it on a Cartesian coördinate system or that of a Polar coördinate system. These can be used to employ algorithms that can be helpful during machine learning to predict the events of each game by player and by team. Subsequently, the outcome of these events can be determined and evidences to known and unknown truths about basketball can be provided. The learning process requires the collection of previous game events by player, team and year; the organization and representation of this information in the most appealing ways to determine the existing patterns (data mining), and consequently the algorithms to be used for the machine learning experiences. However, the data mining method used remains one of the most important decisions for the success of this method. Data analysis by visualization and the analysis of data variations as continuous events rather than discrete ones are especially useful for this process, as it simultaneously interprets various data and provides evidences to this interpretation.

Keywords: data mining, machine learning, basketball, games

I. Introduction

Sports games like basketball necessitate the occurrence of several events during its course whose frequency, time of occurrence and duration vary as affected by other factors. These variations, together with the factors promoting them, form the basis of the intrigue and suspense of sports fanatics before

and during each game. In such situations, many qualifications arise as to the kinds of game events expected. These qualifications may be useful in making better predictions about the types of events that are likely to occur and their outcomes, and this may aid in adapting the structure of the game for any desired outcome. Interestingly, such predictions may not be built on mere hope of the wanted event or event outcome, but on a series of other factors, one of which is the learned structure and/or course of events in more or less similar circumstances of the past. These past structure or behavior are called experiences and are irrefutable thus reliable for the learning process.

Accurately learning from previous experiences is a major challenge. Besides the number of experiences that need to be considered, several other variables exist. One of these is the position on the court at which the event occurs. A 94' x 50' basketball court includes a wide range of positions which can be determining factors of the type of occurrences that may happen during a game. The accuracy in determining the type of event and its outcome on the court is proportional to the number of positions studied on this 94' x 50' court. There is thus the need to study the variations of the experiences on the widest distribution of sectors and subsectors. However, not only do these variations need to the studied, but also the changes that occur as the area of each sector/subsector is increased or decreased. Analyzing these changes is especially useful in predicting the events that are most likely to occur on those "less active" parts of the court (if any exist) on which the available information on events occurring in such area is less significant due to the little amount of data in such areas.

Obtaining knowledge for the learning process can be done in the process called data mining. Data mining is an interdisciplinary branch that encompasses techniques such as machine learning, pattern recognition, statistics, databases and visualization, all in order to provide answers to obtaining information from large databases [2].

The data mining approach used in this research seeks to:

- analyze basketball events at different sectors of the courts;
- analyze how changes in the area of a given sector of a basketball court affect the probability of each event occurring at that sector;
- analyze the pattern formation process as variables and/or sector areas are altered; and
- analyze the outcomes of events occurring at a particular sector of a basketball court.

The method of analysis involves a visual interpretation of the data in order to generate complex relationships in multidimensional data. Programming graphic tools allow both the interpretation of the data and the presentation of this data in appealing format, namely graphs.

II. Related Works

The NBA has used Advanced Scout software to discover patterns in game data [1]. This software uses previous game data to establish these patterns. Several other professional basketball associations are now investing in data mining due to the large amounts of money involved in sports [2]. The differences exist in the data mining techniques used to set up patterns in the data.

Regardless of the technique used, basketball data mining usually requires the setting up of different area of the court to be studied. In [2], the court was divided into eleven positions, and an evaluation of a given game event (shot, rebound) is done within each of the divisions. A basketball court exists as a unit over which variations occur on the range of its area rather than on the subdivision of this area. Therefore, the patterns existing in the game data tend to be over continuous area surfaces rather than on discrete surface areas. In this paper, we study court patterns in game data over adjustable areas on the court and the variation in these patterns as the studied area and other variables are changed. The study is done using visual aids that display the variations as having a continuous effect throughout the court.

III. Procedure and Results

Basketball data collected for over 3,000 games is organized and stored in a database. The database therefore contains information about each event (such as shots, rebounds) that occurred during any of the over 3,000 games. The various game events are grouped by the year and the basketball season during which they occurred; the teams that were involved during the event and; the players that were on the court and, directly or indirectly, involved in the game event. Where needed, the coördinates at which the game event occurred are recorded. For this exercise, the main types of events considered are attempted shots.

Providing this break-down structure to the database will later help in varying the types and number of queries applied to the database. Querying the database will enable a better analysis of the data by allowing us to properly define the variables that might affect the outcome of an event and of a game. Subsequently, the changes caused by such variable during an event or during a game can be studied. This promotes a better understanding of patterns and pattern variations with game event variables. From this structure of the database, three game event variables can be identified: season, team and player.

The database is used to create a graphical user interface with a graph that can depict all shots under these three variables: season, team and player. The graph shows shots taken either in a season, by a team, by a player or a combination of any of these three variables. The XY coördinate graph represents the opponent team's side of the basketball court whose width equals 50' and height 47'. Only a half court is considered because most of the shots occurred on this portion of the shot, and hence analysis made on this sector will be more significant for these studies (Figure 1). The XY plot area is divided into sub areas called grids, which are squared-shaped. Grids of equal areas thus occupy the entire plot representing the basketball court. The grid squares are separated by fine (1.0f) vertical and horizontal lines running from one side of the plot to the other. These lines are kept fine, yet visible, in order to keep an accurate scaled measurement of the grid square areas (Figure 1).

A similar graph is made, but this time a color is assigned to each of the grid squares in which at least one shot was attempted. A shot taken in a grid square was either made or missed. This logic is used to

calculate the probability that a shot taken within a grid area is made. A probability is therefore derived for each of these grid squares using the formula:

$$\frac{\sum \text{shots made}}{\sum \text{shots made} + \sum \text{shots missed}}$$

This yields a value between 0 and 1 inclusive for each of the grid squares. A color gradient varying from red to green is used in coloring (Figure 2) each grid square in which a shot occurred. In the coloring schema, green represents a high probability that shots attempted in the given grid square were made, and red represents a small percentage that shots attempted in the given grid square were missed.

Figure 1: Figure showing the area of the basket ball court that is studied

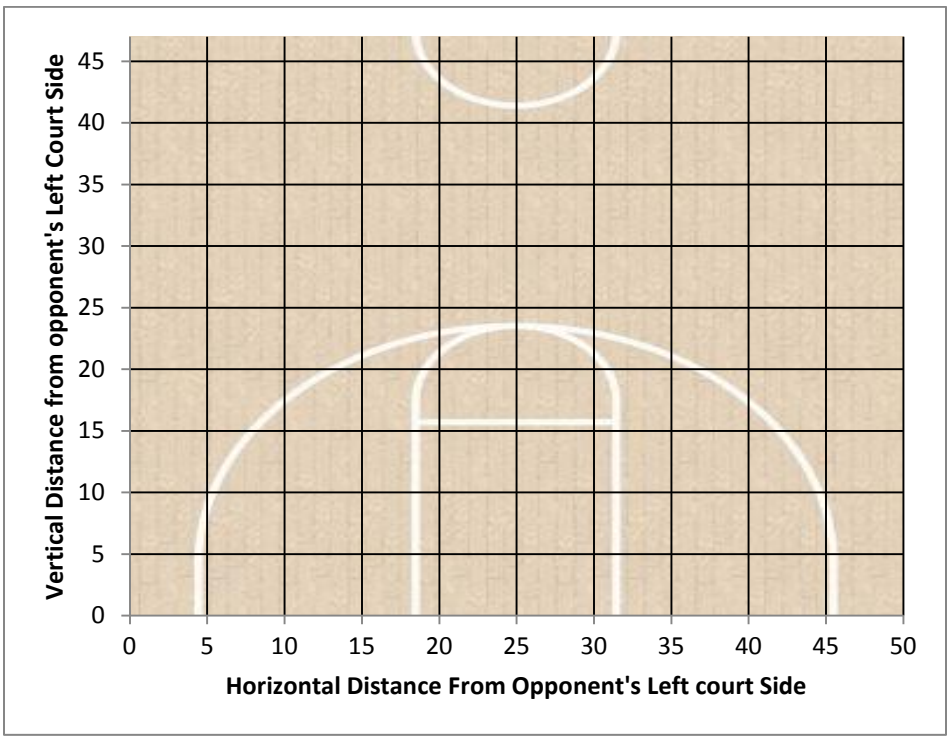

Figure 1: Basketball court divided into resizable square-shaped grids

A slider is added to the plot that is used to vary the area of the grid squares. Moving the slider redraws each of the grid areas by increasing or decreasing the grid area according the slider's direction of motion. Consequently, the probability of shots being made in the new area is recalculated, and a color is re-applied accordingly based on the already defined color gradient scale. The slider is made greatly extensive to allow the variation of the areas over a large range, hence providing room for a more accurate analysis. Although the area of the grids is changeable, the ratio between the areas of any two grids squares is 1 at all times.

Other similar graphs are made to represent information about missed shots. Given that the event following each missed shot is a rebound, information about the type of rebound is derived from the

database. This is done by identifying the player and team that gets the possession of the ball after a shot is missed. Four types of rebounds can be identified: offensive rebounds (OffReb), defensive rebounds (DefReb), offensive team rebounds (OffTmReb) and defensive team rebounds (DefTmReb). Heat maps are created for each of the rebounds on the new graphs to depict the probability of each type of rebound over each grid area where a shot was taken and missed. Four new graphs are thus formed with the same color gradient scale (Figure 2). The color assigned to the grid square areas depends on the value obtained from the following four formulas:

1) $$\frac{\sum \text{OffReb}}{\sum \text{OffReb} + \sum \text{DefReb}}$$

2) $$\frac{\sum \text{DefReb}}{\sum \text{OffReb} + \sum \text{DefReb}}$$

3) $$\frac{\sum \text{OffReb} + \sum \text{OffTeamReb}}{\sum \text{OffReb} + \sum \text{DefReb} + \sum \text{OffTmReb} + \sum \text{DefTmReb}}$$

4) $$\frac{\sum \text{DefReb} + \sum \text{DefTmReb}}{\sum \text{OffReb} + \sum \text{DefReb} + \sum \text{OffTmReb} + \sum \text{DefTmReb}}$$

The first two formulas give the probability of an offensive and defensive player rebound respectively over the total number of player rebounds in a given grid area. The last two formulas are used to obtain the probability of all offensive and defensive rebounds (both player and team rebounds) respectively over the total number of player and team rebounds in a given grid area. As with the previous graphs, a slider is added that varies the size of the grid squares and repaints the new grid squares accordingly. Again, this helps in observing variances in the game events and their outcomes as the grid area change.

Figure 2: Color gradient scale

For each of the graphs, data plots can be queried from the database to depict variations in patterns as the query conditions are changed. Querying is done in the user interface using a drop down menu to select a year, team and/or player data. A query is triggered each time a selection is made using the drop down menu. Because of the necessitated large database size, and the need to be able to visually analyze the changes occurring as a result of the query parameters changing, the database is fully optimized to enhance sorting, filtering and other query functions.

For more sector variations and to analyze the data from a different perspective, corresponding polar graphs are created. The angular coördinates are derived from the (*x, y*) coördinates. Angles measured are from the vertical line through the opponent team's hoop. Angles θ are measured clockwise and counter clockwise, with negative values assigned to all angles to the left of the hoop. Radii are measured from the opponent's hoop at (25, 5.25) to the (*x, y*) position of the game event. Polar coördinates are

therefore in the form (*r, ϑ*) for points on the right side of the hoop (where *x* < 25 or *x* = 25), and (*r, -ϑ*) for those on the left side (where *x* > 25). Because the opponent's hoop is set as the based of the measurement of the angular distances, the values of *r* and *ϑ* are computed different. These are calculated as follows:

$$r = \begin{cases} \sqrt{(25-x)^2 + (y-5.25)^2}, & x < 25, \quad y > 5.25 \\ y, & x = 25, \quad y > 5.25 \\ \sqrt{(x-25)^2 + (y-5.25)^2}, & x > 25, \quad y > 5.25 \\ 0, & x = 25, \quad y = 5.25 \\ \sqrt{(25-x)^2 + y^2}, & x < 25, \quad y < 5.25 \\ y, & x = 25, \quad y < 5.25 \\ \sqrt{(x-25)^2 + y^2}, & x > 25, \quad y < 5.25 \end{cases}$$

$$\theta = \begin{cases} \tan^{-1}\dfrac{25-x}{y-5.25}, & x < 25, \quad y > 5.25 \\ 0, & x = 25, \quad y > 5.25 \\ \tan^{-1}\dfrac{x-25}{y-5.25}, & x > 25, \quad y > 5.25 \\ 0, & x = 25, \quad y = 5.25 \\ 180 - \tan^{-1}\dfrac{25-x}{y}, & x < 25, \quad y < 5.25 \\ \pi, & x = 25, \quad y < 5.25 \\ 180 - \tan^{-1}\dfrac{x-25}{y}, & x > 25, \quad y < 5.25 \end{cases}$$

Where needed, a color is assigned to each polar sector. In this case, two sliders are created: one that varies the angular division of the polar graph and another that varies the space between the polar concentric circles.

Like with the XY plots, a probability rebound rate for all missed shots within all angular and radial sectors where a missed shot occurred is calculated and a color assigned to such sector accordingly. With the sector areas resizable using either/both sliders, the corresponding percentages can be re-calculated upon changes in the area of any given sector, and the appropriate color applied to that sector. The variation of the sector sizes allows the evaluation of each event over 360 degrees angular range and over distances ranges of 25' on either sides of the basketball hoop.

The representation of data on coordinate and polar graphs that are divided into resizable sub sectors allows the observance of the process by which patterns are formed within the plot area representing the basketball court. Knowing the process by which it is formed is helpful in generating these same patterns using alternate means. This process is as important as the patterns itself, for it eases the

derivation of more patterns (either mathematically or by other means) under different conditions. These derived patterns can serve as more learning experience for the learning process.

IV. Conclusion

Data analysis through data mining requires the use of effective methods that studies not only the patterns in the data inputs, but also analyses how these patterns change as some variables are introduced in to the system and eventually altered. Like data, the greater the number of variables (and the alteration of each of these variables) during data analyses, the more efficient is the data mining process. These allow the studying of existing patterns in the data as well as evaluation of the patterns formation process in the data. These studies show that changes occurring per unit time of a basketball game exist as a continuous variable throughout the court.

These different interfaces allowing the study of patterns and relationships in the data provide information according to all the variables considered. This information can be converted into knowledge about historical patterns and future trends in the machine learning process. [3]

V. Future works

Because game outcomes may not be solely dependent on season, team and year, other variables need to be studied. Like the variables used in the experiments, new variables need to be similarly incorporated in the process to generate complex queries that will enhance the accuracy of the predictions. The more variables are considered, the more accurate will the discovered patterns be and the better the outcomes will be for the proceeding stages; and the more complex the queries and the greater the number of queries being processed, the more powerful will the system be.[3]

Having extensively studied game data and patterns observed for different combinations of factors variations, these patterns will be used in machine learning algorithms. Information obtained from the data mining process will be used as input during machine learning for training purposes. Machine learning will then be completed by getting some experience through the observation of patterns and relationships in the inputted data. Successfully applying these processes can yield a machine capable of predicting the events of basketball games as well as the outcome of each event.

VI. Acknowledgements

We thank the Information Sciences Institute at the University of Southern California for providing internet services and an office space.

Thanks to all the members of the Computational Behavior Group and Aaron Henehan for their assistance.

VII. References

[1] Bhandari Inderpal, Edward Colet, Jennifer Parker, Zachary Pines, Rajiv Pratap, and Krishnakumar Ramanujam. Advanced Scout: data mining and knowledge discovery in NBA data. IBM SYSTEMS JOURNAL, VOL 40, NO 3: Ibm T.J. Watson Research Center, 1996. Print.

[2] Beckler Matthew, Hongfei Wang, and Michael Papamichael. NBA Oracle. Pittsburgh: Carnegie Mellon University, 2001. Print.

[3] Frand, Jason. "Data Mining: What is Data Mining?" Jason Frand. UCLA, 7 Aug. 2009. Web. 31 Oct. 2011. <www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm>.