Data mining in sports in the past few years

There has been immense development in the field of sports data mining in the past few years. Starting from the sports enthusiasts who have been trying to better their predictions than their peers, to the novel tools and technologies being developed to enhance personal performance of individual players as well as the overall performance of a team. Prior to the advent of concepts of data mining, all the major sporting agencies emphasized on human expertise. Relying wholly on domain experts was found to be unproductive as time passed by and data grew in scope. With this thought in mind, a quest for statisticians began who would develop more efficient metrics for performance and come up with effective decision making criteria. This was followed by mining the worthful knowledge using the concepts of data mining. Since, the sports domain is so huge, enormous sports data such as statistics, records etc exist. This data comes from the individual performance of players, the championships that a team has played and won, coaching/managerial decisions made in the past and possibly some other game-based events.

This voluminous data, if wisely used, can be of great advantage to any organization by giving it an edge over its peers. The knowledge acquired from the data can be applied to the organization as a whole. Data mining can be used by the players to improve their individual game performance by making use of techniques such as video analysis and by scouts to search and recruit talented bunch of players by exploiting the statistical analysis and projection techniques to maximize throughput. Data mining has thus found its root in the field of sports where the coaches and managers can make decisions and strategies on the basis of important patterns and knowledge extracted from sports data. Since, in today’s competitive sports environment a lot of money has been put on stake, a single decision can prove fatal or fruitful for a sports organization, thus putting it in a much lower or higher position respectively. In such a Data mining has become an integral part of the sports world. Thus, properly exploiting the data mining techniques can lead to better performance by examining the player-situation combination, discovering personal contributions and also by exploiting the patterns which relate to the tendencies of the opponents, statistics of their players, their flaws etc.

DIKW Framework

Data mining is the process of finding or digging out hidden trends and patterns, based on which new data and knowledge can be found out. The data sources could either be structured in the form of databases or unstructured in the form multimedia sources. [1] Data mining concepts are deeply embedded in the field of Knowledge Management. It’s a known fact that before data could be used as knowledge, the intermediate levels should be examined with reference to the Data- Information- Knowledge- Wisdom hierarchy. [1] DIKW hierarchy is a widely accepted concept in knowledge management and each level builds over the top of the previous one in the hierarchy. DIKW framework is responsible for differentiating between data and knowledge and setting boundaries between data, information and knowledge. When applied to the sports domain, some concepts and techniques operate at the data level such as data collection, data mining). On the other hand, certain techniques and algorithms work at the knowledge end, which includes simulations.

2. History

Though there could be numerous relationships between sports and the corresponding sports data, some researchers believe that it can be broadly classified into five classes/levels as shown below in the table.

Level

Relationship

One

Relationship does not exist

Two

Domain experts use their instincts to make predictions

Three

Domain experts predict on the basis of historical data

Four

Statistics are used to make the important decisions

Five

Data Mining is used to make important decisions

Talking more about the relation between sports and the sports related data, we can define what each level actually mean.

Level1: At this level, there is no relationship that exists between sports and its sport-based data. Under this category we have organizations which merely play the sport, collect the personal/team data, but do nothing with it.

Level2: At this level, domain experts play a crucial role in predicting the outcomes of the game or how a particular player/team would play the game, solely based on experience. The decisions so formed by these experts are purely based on instinct, which could include decisions like making a sudden change in the field, moving a player from one place to another as in cricket, or making a substitution as we see in the game of soccer. Such decisions are not based on any prior data, but just gut feeling.

Level3: At this level, historical data that has been collected is the basis on which the domain experts make their decisions. Examples of such decisions would be electing the players by exercising the player-situation combination. A typical example would be from the game of cricket where a person who has performed against certain teams in the past would be played whenever there is game between those two teams. Those decisions would be given higher weightage which would be more fruitful in terms of results based on the historical data.

Level4: At this level, the decisions making process changes a bit and statistics are introduced in the process. Statistics could be used in terms of frequency of certain events which led to better performance or better results. They could be seen in terms of a not so trivial mechanism which would ascribe different players a score or credit on the basis of their individual efforts to achieve a particular milestone. At this level, novel mechanisms of estimating the performance could also be introduced.

Level5: The final level in the hierarchy is the introduction of the data mining. Making decisions on the basis of statistics has been popular for a long time now. But, at the same time statistics do not explain the relationships between patterns and random noise which is done by data mining. This type o relationship can be coupled with the decisions of the domain experts or could be used independently, in which case the decision would be unbiased. The reason behind this is that humans tend to make decisions which are biased towards a particular player but, by removing human intervention in the decision making process and relying completely on the data mining techniques, we can be sure that the decisions would not be biased and thus resulting into better efficiency.

It’s a fact that introduction of statistics has improved the decision making process a lot, but at the same time statistics may be misleading. This is due to the fact that statistics can come from either an imprecise measurement of performance or an over-emphasis of particular statistics by the sports community. [1] An example of this would be that players might have a good individual record or statistics but at the same time don’t contribute immensely to the team performance. Thus, data mining techniques are being adopted by more and more sports organizations these days in order to be at the top.

3. Evolution of data mining techniques in sports

Baseball

The introduction of data mining techniques in the sports arena was not a day’s effort. It began slowly but eventually got adopted by almost all the big sporting nations and organizations in the world today. The game of baseball saw the transition from the statistics being used to data mining techniques being adopted to make important decisions. In 1977, Bill James started writing papers named Bill James Baseball Abstracts. Through his abstracts he exposed the shortcomings of the existing baseball performance metrics and posted his maverick ranking formulae and new statistical performance measures which he named Sabermetrics. [1]

The readers of these abstracts were quite impressed with it but since these methods had not been practically implemented until then, people were apprehensive to incorporate them into the existing system. But, eventually some sabermetricians started exercising these changes and saw some amazing results which landed them into a better position when compared to their peers. At this point in time, despite the success of sabermetrics amongst fans, sports organizations were still unsure if they wanted to incorporate it completely or not, because of the fact that the traditional methods were deep-rooted. [2]

In the early 2000s, Billy Beane, who was the manager of the Oakland A’s baseball team adopted data mining techniques in order to improve the team performance. This led to a period of success for the A’s which entered into the playoffs or playoff contention for five consecutive years.[2] The Boston Red Sox got benefitted from the use of data mining techniques in a similar manner but on a bigger scale as they went on winning the World championship in 2004 and 2007. This could be seen as a phenomenal change in the history of baseball as this team hadn’t won a single world championship in a span of almost 86 years.

The concept of sabermetrics was further exploited by the Bill Beane to build a much more competitive team by selecting the players in the draft as opposed to orthodox manner of doing it. This was then extended to bind the players who were neglected by the other teams to a long term contract. All that mattered was to pick the right set of players which was made a lot easier by sabermetrics.

Soccer

The game of soccer has been a real entertainment source for fans all around the world. The game itself has the biggest fan following, let alone the fan following of individual players. Soccer lacks the grandness that baseball has in terms of statistics. Soccer lacked this grandness of statistics because it was tough to evaluate player activity on the field and quantifying the same added more to the trouble. Billy Beane applied his knowledge and experience to the game of soccer and tried to introduce some sabermetric statistics such as number of touches (frequency of a player being in play), shot creation (if a player acts a participant in a shoot or shoots by himself), ball retention (measure of offensive turnovers), and balls won per 90 minutes of play. [3]

The contribution of Beane in the introduction of statistics and usage of data mining techniques in soccer, to develop an effective strategy for the team selection and improving the game performance was extended by Prof. Anatlov Zelentsov. He fashioned computer programs to not only select the players for team Dynamo, but also to analyze the games that were played by these players.[1] Dynamo used this strategy to win the UEFA cup in 1975 and 1986. Players who were picked up to play for Dynamo were put to different tests including nerve, endurance, memory, reaction and coordination tests. [1]

Once this was done, data mining techniques began to be incorporated to exploit the data that was collected about the players and the games that they played to come up with certain patterns which could explain the flaws and predict the outcome of similar events over a series of future games.

Cricket

Cricket is another sport which is very rich in terms of data. But, unfortunately it does not reflect the true performance measures. Wisden Almanack is one such place where cricket specific data can be found for games which were played as early as 1864. [4] The major issue with this data store is that the data is has been recorded as historical scorecards instead of a collection of player-specific statistics. [5]

John Buchanan, coach for cricket Australia for almost 8 years had a different view of statistics. He advocated the introduction of measures such as pitch conditions, rule changes and equipment changes. He also supported the introduction of those performance measures which represent individual player performance. The next step was pruning the metrics so found so as to focus upon those measures which contributed most to winning as well as assess those statistics to come up with some trends or tendencies which can be exploited. [6]

The new information collected on the basis of statistics can be used to model team performance by comparing the results of keeping a player in the team vs. keeping that player out of the team. The newly gained results can in turn be used to decide if a player is an asset or liability and if a replacement player is worth including in the team as compared to an existing one. [7]

Data mining and knowledge management tools are now being used in cricket to a great extent. When the cricket data was analyzed for one-day and test cricket matches, it was observed that a combination of left and right handed batsmen was an asset to winning. Another criterion that contributed highly to winning was high runs-to-over ratio. [8] These important factors were a result of studying the cricket data patterns and relating them to the no. of wins that a team has experienced. These factors could also be used to predict how well a team would perform against another team in a big tournament. Thus, data mining has proved to be quite effective in predicting a cricket match’s result.

Data Sources

Sports data has seen a revolutionary change in the recent past. In the early days, the data was just recorded and stored simply to keep track of it or for historical purposes. After a span of numerous years, this data began to be explored and looked into by sports analysts which believed that interesting knowledge could be retrieved from the data. This led to a transformation of data stored to meaningful data with contained some patterns, trends or tendencies which could be exploited. This was followed by sports data being stored in highly accessible and searchable form. Data for sports comes from diverse sources.

Professional Societies

There are copious professional societies which share sports data amongst members and also maintain sport related journals and articles. These societies gather, appraise, stock and distribute sporting data while performing further research.

The Society for American Baseball Research (SABR)

This society was formed in Baseball’s Hall of Fame Library in August of 1971. [9] The main concern of this society was to enhance the research in the field of baseball and create a depository of the important baseball data which was not captured by the box scores. SABR research focuses upon individual players or accumulated history of a league. In 1974, SABR founded a committee which came to be known as Statistical Analysis committee (SAC). The research that focused upon evaluating performance data came to be known as sabermetrics which was started just when SAC was formed. The main motive behind the operation of this committee is to study ancient and modern baseball analytically. [10]

Association for Professional Basketball Research (APBR)

This society was formed in 1997 to promote the history and game of basketball and to analyze the statistics of the game in an objective manner. [11] APBR’s main focus was on NBA statistics but it also contained data from other leagues. [12] Just like sabermetrics, APBR developed APBRmetrics to develop better measurement and statistical tools to carry out comparisons. During 1990s, Dean Oliver along with APBR performed further investigation of possession and team related statistics which made APBR as the prime source for quantitative basketball research.

Sports Related Associations

Apart from professional societies related to the sports field, sports related associations also exist which focus on gathering and distributing information to its members. These societies are a bit different from professional societies in the manner that they work. These associations do not adhere to a specific sport, but focus upon improvement of the existing system and techniques as well as archiving the gathered data for future use. `

The International Association on Computer Science in Sport (IACSS)

This association was formed in 1997 with the focus of improving cooperation amongst research groups which are trying to apply the computer science technologies in the sports field. [13] This association shares the research work done by their members by distributing newsletters and journals.

The International Association for Sports Information (IASI)

This association was formed in 1960 with the main focus to standardize and archive world’s sports libraries. [14] This association could be seen as a network of librarians, sport experts and document depositories. Information sharing takes place every three years through newsletters.

Special Interest Sources

Apart from the professional societies and associations, there also exist special sources that gather and perform an analysis of the sports related data. These sources usually share statistics and new data in terms of player biographies, awards won and their personal records.

Baseball

Research in Sports Statistics

After the initial step of gathering the data sports specific data, it is important to look for knowledge which can be derived from this data. Various kinds of statistical analysis can be performed to evaluate a player's performance, balance of a team, the underlying flaws in a team etc. Copious statistics across several game events have been observed in the field of sports. It was not until the latter part of the last century that these statistics were found to be inefficient when compared to the new ones that sports experts like Bill James and Dean Oliver created.

Bill James was made a revolutionary change by introducing a performance metric named sabermetrics, in the game of baseball. Billy Beane, general manager of the Oakland A's adopted sabermetrics in 2002 for drafting players and forming a competitive team.

Baseball Research

Building Blocks

Statistics alone have been used as a means to measure the performance of a player in the past, instead of being the starting phase of a process, the next step of which is to extract useful knowledge. Considering an example of the game of baseball, “ hits” was the most famous statistic. Researchers found that hits does not take into concern other means such as getting on-base and these measures were not even a part of a player’s batting average or the total hits. It was due to this reason, that a new statistic was created which came to be known as On-Base Percentage (OBP). This statistic gave a measure of how often a player gets to the base. Another important statistic that was created was slugging percentage. Taking this statistic into account, the total number of bases reached by a player was divided by the total no. of at-bats and rewards players known for hitting more doubles, triples or home runs instead of just singles. On the other hand, hits considered doubles, triples and home runs to be the same as singles. Another statistic that can be derived from OBP and slugging percentage is On Base Plus slugging statistic (OPS). OPS can be seen as a summation of the previous two statistics and gives a much better measure of the capability of a player to get to the base and hit with power.

Runs Created

Bill in his third abstract supported his theory of measuring the player performance on the basis of scoring more and more runs instead of relying on the prevalent statistic of batting average. [15] He in turn gave a formula for Runs Created which is:

Runs Created= ((Hits+Walks)*âˆ‘ Bases)/ (At-Bats+Walks) [15] 1982

This was by far the most efficient measure of performance since it reflects a team’s ability to score runs by getting to base through the parameters at-bats, hits and walks. The model that James built was then tested by making use of historical baseball data and it was found to be better at making predictions than others. [2] The formula was efficient in the sense that team with higher runs created wins and not the one that has higher batting average. As time passed by, further refinement of the same formula was seen and better variants of the same were developed. Further analysis showed that the players who do not start but come in later have almost 80% of the offensive capability of the starter. This statistic varies for catchers at 85% and first batsmen at 75% the starter's ability.

Win Shares

Next to follow the runs created formula was the criterion of Win Shares. As per this criterion, players were attributed a part of the game winning on the basis of their performance in terms of offensive and defensive input. The whole idea behind win shares is to give credit to players based on their performance which leads the team to a victory. This concept is still being refined but gives a much better perspective of the game in terms of improved performance.

Pitching Measures

Another important statistic that was developed for improving the efficiency and performance of pitchers was the Earned Run Average (ERA). ERA takes into consideration 9 innings to measure performance in terms of number of earned runs and runs which come as a result of hits. The formula that was created is given below:

ERA= (Earned Runs Allowed*9)/ IP where IP stands for the number of innings pitched. [1]

Tools and System for Sports Data Analysis

Due to the advancement of sports to such a big level, data mining and knowledge management tools have gained popularity in the recent past. Those people, who adopted the data mining and knowledge management tools early, got highly benefitted. This resulted into the evolution of advanced measurement techniques. There are a variety of tools available, few of which are described in the section below.

Advanced Scout

Advanced scout is data mining and knowledge management tool which was created by IBM in the latter part of 1990 for NBA. The main task of this tool is to find interesting patterns amongst the NBA game data and benefit the coaches and other scouts by providing a deeper knowledge.

This tool was developed in such a manner that when the game is on, it gathers structured game-related statistics and unstructured multimedia footage as well. This tool has proved to be of immense help to the coaches as well as players in the sense that they can train themselves for an upcoming series or tournament by studying the opponent’s flaws and tendencies using the video footage. [17] This process is very popular in almost every other sport these days since a better strategy can be formed to tackle different situations and produce better results. In principle, this tool follows a sequence of three steps to perform its operation. First, the multimedia part compiles game-time footage and as the second step the content that has been collected is checked for errors. Finally, the footage is segmented into a series of time-stamped events such as shots, rebounds and steals. [18] The first phase of the process which processes the data and checks for errors is a rule-based series of procedures which checks if the data is consistent and accurate. The error checking phase prunes the improperly tagged events and also looks for certain events which could be missing. In certain cases, rule based approach is not suitable and is unable to identify important elements which in turn could be identified by a domain expert. The domain expert can also tag events in the footage by himself.

This tool has a knowledge management element as well which is known as attribute Focusing. As the name suggests, a particular attribute is focused upon and is measured using the complete data that was gathered. The results could then be exhibited in textual as well as graphical descriptions of the abnormal subsets. The subsets which show clear distinction as compared to others are then subjected to more analysis.

Sports Vis

Sports Vis is another data mining tool for sports (baseball), which can be used to find interesting patterns in the collected data. This tool exploits these patterns graphically. The way it works is that a user can see tons of data over a specified time-period. [19] This data can be highly flexible in the sense that a user could select total runs that were scored by a particular team over a specified time or individual statistics such as the total runs that were given away by a pitcher in certain number of games. The main advantage of using this graphical representation of data is that it could help discover some trends or certain issues like injuries. The figure below shows graphical data in terms of runs that a professional pitcher gave over 32 games.

Figure: Runs given by a pitcher in 32 games [19]

Scouting Tools

In the early days, in order to record a player’s performance, scouts had to do a lot of manual work. As technology advanced, certain tools were developed which aided the capturing of player performance which could then be used by scouts. The statistics could be filled in even when the game was going on and information regarding the whole game or personal attributes could be shared or distributed.

Digital Scout

Digital Scout is a software program which can be used by users including sports-fans and sporting organizations to gather game-related statistics and perform an analysis of the same. Digital Scout has the advantage that it can be employed for any kind of sport which involves some sort of statistical record keeping. Using this tool, users can also take a print-out of results of a game and can even focus on a particular attribute to generate reports. This tool has a high utility and has been adopted by various baseball teams.

Inside Edge

This tool was developed in 1984 by randy Istre and Jay Donchetz in an attempt to provide pitch charting and hitting zone statistics for not only professional but college teams playing the game of baseball.[20] This tool became famous within no time since it provides a simple scouting reports in the form of textual and graphical representations and also provides descriptions about the strength, weakness, flaws and tendencies which can be exploited. These reports are supported by data which can be examined by the users. The figure below shows a spray chart of Rafael Furcal of the Atlanta Braves.

Figure: Spray chart for Rafael Furcal [21]

This chart can be studied by the opposing team and it can be observed that the density of the infield shots towards second base in much higher than other parts. Thus, studying such patterns or tendencies the second baseman can expect a large number of ground-hit shots to be directed towards him. Thus, the opposing teams can prepare better keeping in mind how a player plays the game. Such tendencies could be studied, understood and exploited for producing better results.

Another example of a report produced by Inside Edge is the Pitcher Postgame report which shows the increment in the pitching speed of the pitches as game advances. It also shows the pitch effectiveness. Studying this graphical representation, pitchers can have a better idea of their performance in the strike zone. It can also be used to better understand the effectiveness of the opposing pitchers, so as to mould the game accordingly.

The figure below shows one such pitcher postgame report.

Figure: Pitcher postgame report for Bartolo Colon [21]

Predictive Modeling for Sports

As the name suggests, predictive modeling is the process of creating a statistical model which can be used to make predictions for the time to come on the basis of given input data. Its main purpose is to forecast probabilities and trends for the given input. Predictive modeling follows a series of steps, first one being the collection of data for pertinent predictors. This is followed by developing a statistical model which would be used to make predictions. Next comes the most important part of making the prediction and then finally the model so formed is validated when new data is fetched. There are various techniques which can be used for predictive modeling, the prominent ones being simulation and machine learning. Simulation techniques for example BBall in the game of basketball can are widely being used to model a complete season. Using this, favorable substitution patterns can be derived. The question that arises here is that what if the there are certain situations which were not foreseen? The answer to this question is that additional simulations are then carried out to evaluate new types of actions. Apart from this technique, machine learning is widely used to dig out hidden data patterns.

Statistical Simulation

In statistical simulation we simulate the new sport data while keeping the old data as a reference. Once the data has been constructed, a comparison is then carried out against the actual game play for testing the correctness of the predictions so made. Simulation can be applied to various games like baseball, basketball etc.

Baseball

Baseball is a famous game as far the application of statistical simulation goes. In the game of baseball, simulations can be done to find out effective pinch hitters making use of Markov Chains. This is done by considering matrices of players, inning states specifying top or bottom of the inning, total number of outs and on-base possibilities and multiplying these by substitution matrices using the pinch hitters. [22] Optimal patterns for making substitutions can then be found out on the basis of a given circumstance.

As per a simulation technique that focuses on a particular player and makes use of the historic player data, predictions regarding the total future homeruns can be made by carrying out analysis of frequency distribution of homeruns. In this method, extraordinary events such as record breaking seasons are treated as “ large” events and these event frequencies are then mapped to frequencies of the small events such as individual homeruns. [23] Applying this model to extrapolate it for a particular player, their batting tendencies can be predicted. One such example would be when a player has been hitting the ball out of the park more than usual number of times over a season, then it can be predicted that he will have a high scoring season.

Another example could be the predictions made regarding baseball’s division winners. It uses a Bayesian model which has a couple of stages and is based on a team’s relative strength which is measured in terms of winning percentage, batting average, ERA of a starting pitcher and home ground advantage. [24]

This study showed great accuracy and MLB baseball’s whole 2001 season was simulated using the same technique. It accurately predicted five games out of six with a success rate of 86%.

Basketball’s BBALL

Basketball is another sport which extensively uses simulation techniques to make predictions. One such popular tool in the sport of basketball is BBall. This tool was created to aid NBA coaches, scouts and managers to find out an efficient substitution patterns which would produce the most stimulated wins over an entire season. In the recent past, BBall has been used to find out the consequence of inclusion and exclusion of a player from a team, the consequence of a player being injured etc. It can also be used to magnify the performance of a team in terms of key factors such as rebounds, assists and scoring.

Machine Learning

Another important branch of predictive modeling is machine learning which has been in use for a long time now. This is an alternative of using simulation techniques to predict a game-based event.