The Use of Data Mining for Basketball Matches Outcomes Prediction

Dragan Miljković, Ljubiša Gajić, Aleksandar Kovačević, Zora Konjović
Faculty of Technical Sciences, Novi Sad, Republic of Serbia
dramilj@hotmail.com, ljubisa_gajic@yahoo.com, kocha78@uns.ac.rs, ftn_zora@uns.ac.rs

***Abstract* — Sport result prediction is nowadays very popular among fans around the world, which particularly contributed to the expansion of sports betting. This makes the problem of predicting the results of sporting events, a new and interesting challenge. Consequently systems dealing with this problem are developed every day. This paper presents one such system, which uses data mining techniques in order to predict the outcomes of basketball games in NBA (National Basketball Association) league. The problem of predicting the game result is formalized as a classification problem, where the Naive Bayes method is used. Besides actual result, for each game system calculates the spread, by using multivariate linear regression. The MVC Model 2 pattern based software system is implemented. The system was evaluated on the dataset comprising 778 games from the regular part of the 2009/2010 NBA season and it correctly predicted the winners of about 67% of matches.**

I. INTRODUCTION

Sport result prediction is nowadays very popular among fans around the world, which is particularly contributed to the expansion of sports betting. This is particularly evident for the most popular sports such as football and basketball. Many people have developed various systems with the aim of providing the best possible prediction of the winners of sporting events. The main problems with such systems are that users are often swayed by emotions or the systems do not work with the right set of data. One approach to this prediction problem is the use of data mining techniques, which will be demonstrated in this paper. This paper presents a system which predicts the outcomes and calculates the spread of NBA games. Prediction problem is formalized as a problem of classification, while for the calculation of the spread multivariate linear regression is used. Several classification methods were tested and among them, Naive Bayes method is selected because it provided the best results during testing phase. The system was evaluated on the collection of 778 games from 2009/2010 season and system has correctly predicted the winners of about 67% of matches.

The rest of the paper is organized in five sections describing related work, the problem and solution method, software implementation, system's evaluation and prediction results, and concluding remarks.

II. RELATED WORK

There are many papers dealing with the topic of predicting the results of sports events, and most of them are related to football. A variety of techniques are used, ranging from data mining to statistical methods. In [1] authors present the approach for predicting the results of football matches called FRES (Football Result Expert System). A combination of Bayes inference and rule-based reasoning is used. Each game is represented as a series of tides (flows). Probabilities of some events, like possible player substitute or change of formation, are calculated on the basis of previous data in each tide and are used to fire certain rules to determine the decisions for the next tide. The system was applied to the FIFA World Cup 2002 matches and it correctly predicted the champion and runner-up, and also 6 of 8 participants in the quarter-finals.

A common approach to this subject matter is also the use of artificial neural networks, which can be seen in [2] and [3]. The idea behind these works is the use of a number of statistical data to train the neural network (multi-layer perceptron in most of the cases). The system presented in [3] is evaluated in the international competition in result predictions - Top Tipper [4] and proved to be satisfactory by comparing it with other competitors. Average percentage of correct predictions for ranged from 54.6% for English Football Premier League to 67.5% of Super Rugby.

Naive Bayes classifiers are used to predict the Cy Young Award winners in American baseball in [5]. In the period since 1967 to 2006 this system successfully predicted over 80% of the prize winners.

A form of Elo rating system [6], modified is such way that to allow "home ground advantage", is used in [7] for generating prediction probabilities. The Minimum Message Length [8] is used as a parameter estimation technique.

In the basketball domain, data mining techniques are mostly used for purposes other than the prediction of match outcomes. Authors in [9] describe a system based on the statistical data and fuzzy evaluation which provides a performance evaluation of players and can also predict their performances in the upcoming matches. A solution for supporting the basketball coaches in making tactical/technical decisions during matches and also for pre-game and post-game analysis is presented in [10]. The system is based on the use of apriori algorithm and is tested on data from the Italian first division.

Lately a lot of work is done on the subject of processing and extracting relevant information from the video clips of sports matches. Solutions that address this topic are described in [11] and [12].

978-1-4244-7395-3/10/$26.00 ©2010 IEEE

III. THE PROBLEM AND THE SOLUTION

The problem intended to be solved by the software proposed in this paper is the following one: predicting the outcomes of the NBA matches and calculating the spread for NBA games.

The outcome predicting problem is formalized as a classification problem, where the match outcomes belong to exactly one of the two categories (classes): those in which the host will win and those in which the visiting team will take the victory (unlike some other sports, in basketball there is no draws and teams are forced to play extra time until there a winner is determined).

Spread (which is used only in the sports betting) represents an advantage in the number of points given to one of the teams in order to equalize their chances for victory in the eyes of bookmakers.

Each game is described with record consisting of 141 attributes, which are related to the outcome of the match, and the participating teams. For each team there are 2 groups of attributes.

The first group consists of standard basketball statistics (field goals made, field goals attempted, 3 pointers, free throws, rebounds, blocked shots, fouls, etc). The attributes in the second group are composed of the information about league standings (number of wins and loses, home and away wins, current streak etc). For the first attribute group overall statistics are taken into account, as well as those from home games and those from away matches. Complete statistics attribute set with descriptions can be seen in Table I, while standings attribute set is described in Table II.

Multivariate linear regression is used for calculation of the spread for NBA games where the point difference was used as a target variable, while all other attributes were used as explanatory variables.

TABLE I.
STATISTIC ATTRIBUTE SET

Attribute	Description
FGM	Field goal made per game
FGA	Field goal attempted per game
FGP	Field goal percentage per game
3M	3-pointers made per game
3A	3-pointers attempted per game
3P	3-pointers percentage per game
FTM	Free throws made per game
FTA	Free throws attempted per game
FTP	Free throws percentage per game
OR	Offensive rebounds per game
DR	Defensive rebounds per game
TR	Total rebounds per game
AS	Assists per game
TO	Turnovers per game
ST	Steals per game
BL	Blocks per game
F	Fouls per game
P	Points per game

TABLE II.
STANDINGS ATTRIBUTE SET

Attribute	Description
Won	Total number of wins
Lost	Total number of losses
Pct	Percent of wins
Homewon	Number of games won at home
Homelost	Number of games lost at home
Roadwon	Number of games won away
Roadlost	Number of games lost away
Divwon	Number of games won in division
Divlost	Number of games lost in division
Confwon	Number of games won in conference
Conflost	Number of games lost in conference
Streak	Number of consecutive wins/losses
L10won	Number of wins in last 10 games
L10lost	Number of losses in last 10 games

In order to improve the performance of the classification in the system, feature selection and normalization were used. Experiments were carried out with the variety of classification techniques: decision trees, k nearest neighbors, Naive Bayes and support vector machines. The best results were obtained by the Naive Bayes classifier combined with normalization and feature selection.

IV. SOFTWARE IMPLEMENTATION

For development of a classification and regression models that are used in the system, RapidMiner [15] was selected as a tool. Also, RapidMiner's implementation of naive Bayes method and multivariate linear regression are used in application itself in the form of Java libraries.

All data used in this software are stored in MySQL relational database and accessed using JDBC (Java Database Connectivity) drivers. Database schema can be seen in Figure 1.

The complete system for result prediction was implemented in Java programming language using MVC Model 2 pattern [16] based on the synchronous work of three components: the model, visualization layer and controller. Figure 2 shows these components and how they interact together.

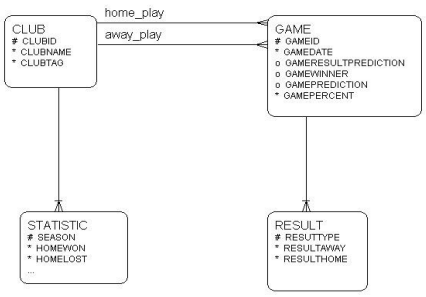

Figure 1. Database schema

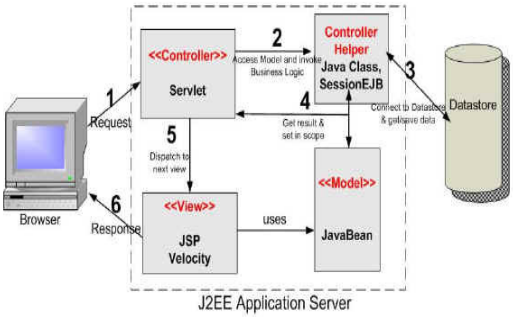

Figure 2. MVC Model 2 pattern

The model component models data and processes. This component interacts with the database and executes operations over data. Entity JavaBeans components, Java classes mapped to a corresponding table in relational database are used as model. For communication with the database there are Controller Helper components, including Session Enterprise JavaBeans classes and DAO (Data Access Object).

The view component (visualization layer) visualizes data provided by the model component. JSP (Java Server Pages) with JSTL (Java Server Pages Standard Tag Library), EL (Expression Language) and JavaScript is used in this software.

The controller manages the visualization layer and the model. All user actions go through this component which makes the decision on which application part will be executed and displayed. The roles of controller in this application perform servlets, more specifically Java Servlet classes.

The activity diagram is displayed on Figure 3.

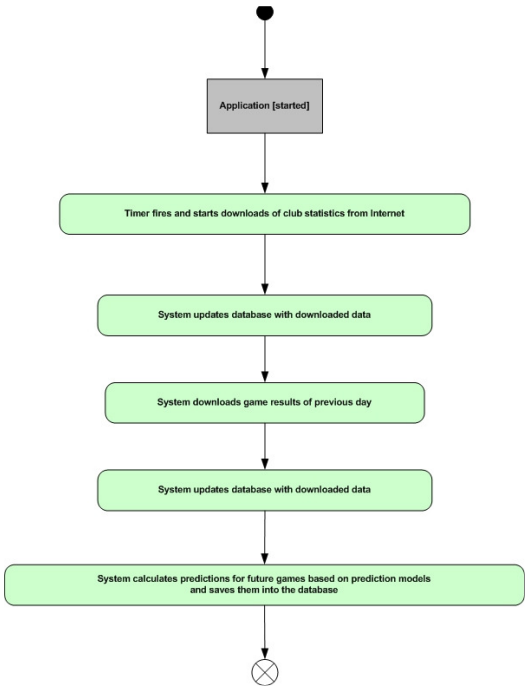

Figure 3. UML Activity diagram

From the Figure 3 one can see that the application operates in the following manner:

- Fresh information about the games that were played previous day, as well as new statistical data for all teams are downloaded.
- Based on these data and prediction models the system calculates the predictions for future games in the form of probabilities (for each team the likelihoods of winning a match are given).

The application allows user to view predictions for the upcoming matches, as well as the results of games already played and previous predictions. Some screenshots can be seen in Figures 4 and 5.

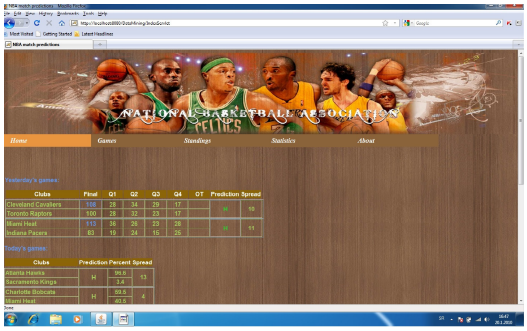

Figure 4. Home page

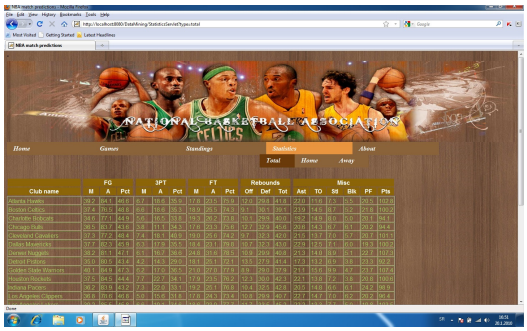

Figure 5. Statistics page

V. EVALUATION AND RESULTS

In order to evaluate the system, the statistical data of games that have already been played during a 2009/2010 season were used. The dataset consists of 778 games. All data were collected from the official NBA website [13] and part of the Yahoo website dedicated to NBA league [14]. One example from the dataset is given in Table III. This example represents the game between Sacramento Kings and Houston Rockets, played at April 12th, 2010 in Sacramento, California.

K fold cross-validation was used for evaluation of the system. In this validation method the dataset is divided into *k* subsets and each of them exactly once used as a test set, and in other *k-1* iterations as a training set. 10 fold cross-validation was used. Accuracy was used as the

performance measure, and it is calculated by the following formula:

$$accuracy = \frac{number\ of\ correct\ \ predictions}{total\ number\ of\ \ predictions} \quad (1)$$

TABLE III. DATASET EXAMPLE

Attribute	Sacramento Kings	Houston Rockets
Won	25	42
Lost	55	39
Pct	0.313	0.519
Homewon	18	23
Homelost	22	18
Roadwon	7	18
Roadlost	33	22
Divwon	5	9
Divlost	10	7
Confwon	16	27
Conflost	34	24
Streak	-1	-1
L10won	1	5
L10lost	9	5
FGM	38.3	37.7
FGA	84.1	84.4
FGP	45.6	44.7
3M	5.9	7.9
3A	16.9	22.4
3P	34.9	35.2
FTM	17.4	19.0
FTA	24.0	24.7
FTP	72.6	77.2
OR	11.9	11.8
DR	30.7	30.1
TR	42.6	42.0
AS	20.5	21.8
TO	14.2	13.8
ST	6.9	7.1
BL	4.5	3.9
F	22.3	20.9
P	100.0	102.4
Winning probabilities	42.8	57.2
Actual result	107	117

The system's achieved accuracy was of 67%, which means that it successfully predicted the outcome of more than two-thirds of the games in that period (a total of 778 games). This can be considered a very good result by comparing it with other systems. For example, sports journalists from CBS had the percentage accuracy of 68.3% in the previous season [17]. The overall results in terms of precision and recall are given in Table IV.

TABLE IV.
CONFUSION MATRIX

		Predictions		Recall
		Home win	Away win	
Actual results	Home win	354	100	77,97%
	Away win	156	168	51,85%
Precision		86,34%	62,69%	

As for the prediction of spread, the system has successfully provided predictions for 78 of 778 matches (~ 10%), which is the expected result, because it is very hard to predict the exact difference in which basketball game ends and the goal of this attempt was to provide only approximate information on the possible difference.

VI. CONCLUSION

This paper presents a system for predicting the outcomes of NBA league matches. The application provides predictions about the winner of the match (home or visiting team) and the value of spread for the match. Naive Bayes classification method and multivariate linear regression were used to perform these tasks. The achieved results are satisfactory and in line with expectations.

Future plans include experimenting with the system by applying it to other sports domains such as various football competitions, American football, hockey etc. and also experiments with some other methods of classification, such as neural networks. .

REFERENCES

[1] Byongho Min, Jinhyuck Kim, Chonghyoun Choe, Hyeonsang Eon, Robert Ian (Bob) McKay, "A Compound Framework for Sports Prediction: The Case Study of Football", *Knowledge-Based Systems*, vol. 21, issue 7, October 2008, pp. 551-562, Elsevier Science Publishers B. V. Amsterdam, The Netherlands

[2] Michael Baulch, "Using Machine Learning to Predict the Results of Sporting Matches", Department of Computer Science and Electrical Engineering, University of Queensland

[3] Alan McCabe, Jarrod Trevathan, "Artificial Intelligence in Sports Prediction", *Proceedings of the Fifth International Conference on Information Technology: New Generations*, 2008, pp. 1194-1197, IEEE Computer Society

[4] http://www.toptipper.com

[5] Lloyd Smith, Bret Lipscomb, Adam Simkins, "Data Mining in Sports: Predicting Cy Young Award Winners", *Journal of Computing Sciences in Colleges,* vol. 22, issue 4, April 2007, pp. 115-121, Consortium for Computing Sciences in Colleges

[6] http://en.wikipedia.org/wiki/Elo_rating_system

[7] Ryan Baird, "Probabilistic Sports Prediction Using Machine Learning Honors Thesis", School of Computer Science and Software Engineering, Monash University, Clayton, Vic.

[8] http://www.csse.monash.edu.au/~lloyd/tildeMML/

[9] Mourad Atlas, Yan-Qing Zhang, "Fuzzy Neural Agents for Online NBA Scouting", *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 58-63, IEEE Computer Society

[10] Guiseppe Polese, Massimiliano Troiano, Genoveffa Tortora, "A Data Mining Based System Supporting Tactical Decisions", *SEKE '02: Proceedings of the 14th international conference on Software engineering and knowledge engineering,* 2002, pp. 681-684, ACM

[11] Sebastian Lühr, Mihai Lazarescu, "A Visual Data Analysis Tool for Sport Player Performance Benchmarking, Comparison and Change Detection", *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, vol. 01, 2007, pp. 289-296, IEEE Computer Society

[12] Lawrence Y. Deng, Yi-Jen Liu, "Semantic Analysis and Video Event Mining in Sports Video", *Proceedings of the 22nd International Conference on Advanced Information Networking and Applications – Workshops*, 2008, pp. 1517-1522, IEEE Computer Society

[13] http://www.nba.com

[14] http://sports.yahoo.com/nba

[15] http://www.rapid-i.com/

[16] Branko Milosavljević, Milan Vidaković, „Java i internet programiranje", 2007, FTN Izdavaštvo, Novi Sad

[17] http://www.cbssports.com/nba/exclus1ives