A Review of Data Mining Techniques for Result Prediction in Sports

Maral Haghighat [1], Hamid Rastegari [2] and Nasim Nourafza [3]

[1] **Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University,
Najafabad, Iran**
maral.haghighat.1366@gmail.com

[2] **Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University,
Najafabad, Iran**
rastegari@iaun.ac.ir

[3] **Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University,
Najafabad, Iran**
n_noorafza@yahoo.com

Abstract

In the current world, sports produce considerable statistical information about each player, team, games, and seasons. Traditional sports science believed science to be owned by experts, coaches, team managers, and analyzers. However, sports organizations have recently realized the abundant science available in their data and sought to take advantage of that science through the use of data mining techniques. Sports data mining assists coaches and managers in result prediction, player performance assessment, player injury prediction, sports talent identification, and game strategy evaluation. The present study reviews previous research on data mining systems to predict sports results and evaluates the advantages and disadvantages of each system.

Keywords: *Sport Matches , Data Mining Techniques, Result Prediction, Prediction Accuracy.*

1. Introduction

As a scientific field, physical education discusses the application of scientific principles and techniques to improve sports performance [1]. Since the relationships between sports results and various data elements are directly affected by several factors such as type of sports, the environment, and the objectives of players, several methods have been suggested to predict the results based on available data. More precisely, while some teams prefer not to use any prediction techniques, others have long depended on either the experience and instincts of the experts (with high error rate) or historical data. Teams seeking more reliable predictions, however, tend to take advantage of statistics in decision-making. The most recently developed, and yet least frequently employed, technique in this field is data mining [2]. Data mining systems aim to assist coaches and sports managers in not only result prediction, but also player performance assessment, player injury prediction, sports talent identification, and game strategy evaluation.

Predicting game results has become widely popular among sports fans, especially soccer and basketball fans, throughout the world. However, the numerous systems proposed for this purpose are challenged by some major problems. For instance, the developers may be influenced by emotions or the systems may not work well with datasets. Hence, data mining techniques have attracted attention [3].

The present research reviewed a number of game result prediction systems based on data mining techniques. It included data collection and feature selection methods in the second and third parts, classification techniques in the fourth part, and the results, advantages and disadvantages of the selected systems in the fifth part. The last part comprises conclusions and suggestions for better future systems.

2. Data Collection

While a great deal of sports data can be effective in data mining, lack of a general dataset forces the researchers to collect the required data, either manually or automatically, from valid sports websites (Table 1).

3. Feature Selection

After data collection and adding new features to the existing ones, the accuracy and speed of predictions will depend on proper manual or automatic selection of the most significant, highly correlated features.

Kahn evaluated primary features and employed the method suggested by Purucker [20] to select five final features for prediction [4]. McCabe and Travathan assessed features

Copyright (c) 2013 Advances in Computer Science: an International Journal. All Rights Reserved.

related to various sports and selected 11 that were common among all sports [6]. Zdravevski and Kulakov used experts' opinions and manually chose 10 features with the greatest effects on the results [7]. Trawinski employed eight feature selection algorithms in the Waikato Environment for Knowledge Analysis (WEKA) and picked five features (out of 15) that have been repeatedly selected by the algorithms [9]. Ivankovic et al. weighted nine features using a neural network. They concluded that defensive rebound and number of assists had the highest and lowest effects on a team's win [15].

Davoodi and Khanteymoori chose a combination of eight symbolic and numeric features for training and test phases. After the encoding of symbolic features into numbers, they normalized the numbers due to their various ranges [13]. Buursma considered an initial dataset containing 10 features. Then, he eliminated one feature at a time and applied the classification algorithm. A feature was completely excluded if its elimination improved the classification accuracy. He selected a total of eight features at this stage. He continued the procedure by adding one feature at a time and applying the classification algorithm. The extra feature was not accepted unless the classification accuracy was increased. Finally, among the eight added features, only one was approved and a final of nine features were selected [16]. Cao used manual methods based on SQL and basketball experts' reports and automatic methods based on SQL reports and Ruby script to select 46 features out of 60 items [18].

4. Classification Techniques

The existing potential to analyze and understand large datasets is far below that to collect and maintain data. Therefore, a new generation of techniques and instruments are being developed to aid humans with intelligent analysis of this high volume of data and to result in critical knowledge [21].

Sports provide huge data about each player, team, game, and season and are thus perfect for testing data mining techniques and instruments [18]. Since experts and statisticians cannot explain relations within data for a single game, data mining techniques are employed to assist the experts or to be used independently in decision making [2]. Consequently, sports teams can gain advantage over their rivals by converting data to applied knowledge through appropriate data extraction and interpretation [22]. Moreover, although various unidentified factors may influence the results, data mining will still be valuable in result prediction [18]. However, it has not been fully exploited in this field. Considering recent research on the use of several data mining techniques including artificial neural networks [4,6,9,15,17,18], decision trees [7,9], Bayesian method [3,7,16,18], logistic regression [7,16,18], support vector machines [18], and fuzzy methods [9] in predicting sports results, we review these techniques in the following sections.

4.1 Artificial Neural Networks (ANN)

ANN are novel computational methods to predict the output of complicated systems through machine learning and representation and application of knowledge. As they mimic a biological neural network, they consist of a number of interconnected neurons (processing elements) in particular layers. Neurons in each layer have weighted connections with neurons in the previous and next layer. An ANN comprises at least one input layer, one output layer, and some hidden layers if necessary. During the learning phase, an ANN processes a training dataset and seeks proper weights for the network to correctly classify all training data (a well-known training algorithm is error back-propagation) [23].

Kahn used a multilayer perceptron neural network (with 10, three, and two nodes in the input, hidden, and output layers, respectively) to predict football results. He used two datasets of the mean statistics of one season and the mean data of each team during the last three weeks of the season. He trained the network by error back-propagation algorithm [4]. Similarly, McCabe and Travathan used a multilayer neural network (with 19-20, 10, and one nodes in the input, hidden, and output layers, respectively) to predict football, soccer, and rugby results. Since they were more concerned about the accuracy of prediction rather than its speed, they trained the network with error back-propagation algorithm instead of the conjugate gradient method. They normalized the output to a value between zero (lose) and one (win) [6].

Davoodi and Khanteymoori designed a multilayer feedforward neural network (with one, two, and one input, hidden, and output layers, respectively) to predict the finish time for each participant and thus the standings. All the nodes in each layer were interconnected with the nodes in the next layer. They trained the network with error back-propagation, back-propagation with momentum, quasi-Newton, Levenberg-Marquardt, conjugate gradient descent method. They sought to minimize the Mean Squared Error (MSE) and found the best model to have eight (equal to the number of features) nodes in the input layer, five in the first hidden layer, seven in the second hidden layer, and one in the output layer [13]. Ivankovic et al. tried to predict basketball results using ANN. They first applied a feedforward neural network with 12 input nodes, one output node, and one hidden layer to assess the effects of various types of shots on game results. They then improved the findings by adding the effects of statistical parameters to a network with nine input nodes, one output node, and one hidden layer. Cross Industry Standard Process for Data Mining was followed throughout the

Copyright (c) 2013 Advances in Computer Science: an International Journal. All Rights Reserved.

study and both networks were trained with error back-propagation algorithm [15].

4.2 Support Vector Machines (SVM)

Generally, an SVM uses non-linear mapping of the training set with high dimensionality. In other words, the algorithm searches for an optimal separating hyperplane which acts as a decision boundary between the two classes. An SVM will find the hyperplane by employing vectors (training dataset) and margins (defined by vectors). Although the training of an SVM takes more time compared to other methods, the algorithm is believed to have high accuracy owing to its high capability in building non-linear, complex decision boundary. It is also less prone to overfitting [24].

Cao employed an SVM, a simple logistic classifier (a combination of algorithms whose core is logistic regression and uses LogitBoost as a simple regression function [25]), naïve Bayes, and a multilayer perceptron neural network to predict basketball results. Due to the existence of two groups, two outputs were considered. The output with greater value was selected as the prediction. He then tested the practicality of his models by a scoring process (a process to test a model in prediction of events not yet occurred) [18].

4.3 Bayesian Method

The Bayesian model is among the most famous supervised classification techniques in machine learning. It is simple and efficient and works well on data with various unrelated features or high levels of noise. Bayesian classifier is a probabilistic prediction model that assumes all features to be conditionally independent from the target variable, i.e. there are some unrelated features in each class. It then predicts the new data according to previous data. The Bayesian algorithm uses Bayes theorem with the following formula [23]:

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[A]} \qquad (1)$$

Bayesian networks are graphical models for inference in the presence of complexity and uncertainty. They are directed acyclic graphs that show random variables (as nodes) and their conditional dependencies [26]. A Bayesian network is easy to under-stand and develop and works even with incomplete data [27].

Miljkovic et al. used a naïve Bayes model to predict the results of NBA games. They categorized game results as win and lose and implemented the model in RapidMiner environment. Every day, they added data about the previous day’s games to the existing data in the system. The system then provided a probabilistic prediction of future games based on the updated dataset [3].

4.4 Decision Trees

Decision trees are powerful and common classification and prediction tools. In contrast to ANN which only provide the final prediction and keep the path hidden in the network, decision trees result in a set of rules that clarifies the final prediction [28]. A decision tree poses questions about data features and classifies the data accordingly. Each question is the subset of a node and each interior node points to a child node for each possible answer to the question. Therefore, the posed questions will form a hierarchy and finally a tree. Classification is performed by making a path from the root (the top node) to a leaf (a childless node) [29].

Zdravevski and Kulakov used classification techniques available in WEKA (including decision trees) to predict the winner of a game. They designed a module to collect data, create new features out of the existing features, select features, and classify data in WEKA [7].

4.5 Fuzzy System

Fuzzy logic is a newly developed technology that contributes to the development of systems requiring advanced and complicated mathematical analyses. While in traditional binary sets variables can take either zero or one, fuzzy logic variables may have a truth value which ranges between zero and one [30]. Fuzzy systems can precisely describe indefinite, irrational phenomena. They work based on IF-THEN rules (continuous membership functions) stored in a knowledgebase. In fact, a fuzzy system transforms human knowledge into a mathematical formula [31].

Trawinski designed a fuzzy model to predict ACB league results. He considered the problem as a binary classification and used a three-phase modeling process including data collection, data preprocessing (missing value correction, feature selection, and data scaling), and implementation of 10 learning algorithms using Knowledge Extraction based on Evolutionary Learning (KEEL) [9].

4.6 Logistic Regression

Logistic regression is a well-known tool for classification problems. Like linear regression, logistic regression depends on linear combination of features which is then mapped to a value between zero and one by the logistic function [32]. Therefore, dependent variables should have a continuous value which is in turn a function of the odds of the events [33]. Logistic regression involves two stages: first, estimating the odds of characteristics of each group, and second, determining cut-off points and categorizing the features accordingly. The coefficients are estimated by maximum likelihood estimation [34]. Logistic regression

Copyright (c) 2013 Advances in Computer Science: an International Journal. All Rights Reserved.

has attracted wide attention due to its simple calculations and interpretation along with reliable results [23].

Buursma selected a set of features and used a number of classification algorithms including simple and logistic regression, Bayesian network, naïve Bayes, and decision tree to predict soccer match results. His predictions had three outputs (i.e. win of the host team, draw, win of the guest team). The odds for these three outputs in each game were calculated and the output with the highest odds was selected [16].

5. Evaluation of the Results

Since the accuracy of all designed models has to be assessed, researchers use a learning and a test dataset for this purpose. This part presents the accuracy of the aforementioned studies and summarizes their advantages and disadvantages in Table 2.

Kahn worked on the 14th and 15th weeks of NFL league in 2003. The accuracy of prediction was higher for the whole season than for the last three weeks (75.0% for the whole season vs. 62.5% in the 14th week and 37.5% in the 15th week) [4]. McCabe and Travathan tested the last three seasons of each league. They used ANN to predict the results of AFL, NRL, EPL, and Super Rugby League and obtained accuracy of 65.1%, 63.2%, 54.6%, and 67.5%, respectively [6]. Miljkovic et al. used k-fold cross-validation to classify the training and test datasets and found an accuracy of 67.0% (correct prediction of two thirds of the games [3].

Zdravevski and Kulakov used training and test datasets including 930 games. They classified the data with 37 algorithms in WEKA and compared the results with those of a reference classifier. The reference classifier applied the Hollinger formula [35] to calculate a rate of A for the host team and a rate of B for the guest team. If A – B + 3 > 0, then the win of the host team was concluded (three was added as the “host advantage”). Otherwise, the guest team was supposed to win. After all analyses, the accuracy of the reference classifier was 5% below that of some other classifiers [7]. Ivankovic et al. considered 75% of the collected data as the training set and the rest as the test set. Following the application of an ANN on test data, they obtained an accuracy of 80.96% [15].

Trawinski evaluated his designed system with 10-fold cross-validation. He performed two sets of tests. First, the results were predicted according to the last three games. This model had six features and he concluded that Clas-Fuzzy-LogitBoost had the best accuracy (82.0%) and Clas-Fuzzy-Chi-RW algorithm had the best standard deviation (0.01). The, the results were predicted based on the team’s current status and statistics during one season. This second model had 15 features. Analysis on the test dataset suggested Clas-Fuzzy-Chi-RW algorithm to have the highest accuracy (71.5%) and Clas-Fuzzy-Ishib-Weighted and Clas-Fuzzy-Ishib-Hybrid to have the best standard deviations (0.063). He finally selected Clas-Fuzzy-Chi-RW algorithm for predictions [9].

Davoodi and Khanteymoori used an ANN with five learning algorithms to predicting the results of horse races. They found conjugate gradient descent (CGD) to be the most appropriate algorithm for predicting the last horse. Back-propagation and back-propagation with momentum were the best algorithms to predict the first horse. The Back-propagation and Levenberg-Marquardt required the longest and shortest training times, respectively. Finally, the back-propagation algorithm had the highest accuracy (77.0%) [13]. Buursma employed ClassificationViaRegression (linear regression), MutiClassClassifier (logistic regression), RotationForest (decision tree), BayesNet, and naïve Bayes to classify data in WEKA and reported the obtained accuracy as 55.05%, 54.98%, 57.00%, 54.55%, and 54.43%, respectively [16]. Cao collected NBA league data during 2006-10 and portioned the collected data using k-fold cross-validation. He observed accuracy of 67.82%, 65.82%, 67.22%, and 66.67% using simple logistic classifier, naïve Bayes, SVM, and ANN, respectively. He then considered data from 2010-11 season as the scoring dataset and calculated the accuracy as 69.97%, 66.25%, 67.70%, and 68.01% by employing the above-mentioned techniques, respectively [18].

6. Conclusion and Suggestions

Considering the popularity of sports in the current world, many organizations disburse large funds to gain better results in sports matches. Therefore, predicting game results has turned into a subject of interest for different sports organizations. Data mining, a widely accepted method to predict and explain events, is an appropriate tool for this purpose. Various data mining techniques such as ANN, decision trees, Bayesian method, logistic regression, SVM, and fuzzy methods have been employed to predict game results in recent years. We evaluated available literature in this regard and detected two major challenges. First, low prediction accuracy highlighted the need for further research to obtain reliable predictions. Second, lack of a general and comprehensive set of statistics forces the researches to collected data from sports websites. Difference in the used datasets prevents the researchers from comparing their results with previous studies and leads to unclear development.

We may suggest a number of solutions to eliminate such challenges. For instance, prediction accuracy can be improved through the use of machine learning and data mining techniques that have not been used in this field but have yielded good results in other fields. Application of

Copyright (c) 2013 Advances in Computer Science: an International Journal. All Rights Reserved.

hybrid algorithms can also boost prediction accuracy. Moreover, including different features such as player performance will contribute to more accurate predictions. On the other hand, a comprehensive dataset can be collected by the help of a group of experts in each sports field. In order to provide the chance for comparisons between different studies, researchers are recommended to collect data from valid leagues (e.g. NBA).

Table 1: Datasets of the studied systems

Researcher (Year)	*DataSet*	*Source*
Kahn [4] (2003)	The first 15 weeks of National Football League, 2003	National Football League official website [5]
McCabe and Travathan [6] (2008)	Australian Football League, National Rugby League, the English Premier League, and Super Rugby	-
Zdravevski and Kulakov [7] (2010)	Two consecutive seasons of the National Basketball Association (NBA) League	Basketball-Reference website [8]
Trawinski [9] (2010)	Asociacion de Clubes de Baloncesto (ACB) Basketball League statistics from 2008-2009 season	ACB official website [10]
Miljkovic et al. [3] (2010)	NBA League, 2009-2010 season	NBA official website [11] and Sports.Yahoo.com [12]
Davoodi and Khanteymoori [13] (2010)	A hundred AQUEDUCT Race Track games during January 1st-29th, Ney York, USA	Equibase website [14]
Ivankovic et al. [15] (2010)	Basketball League of Serbia B, 2005-2006 season until 2009-2010 season	Basketball Federation of Serbia/Basketball Supervisor software
Buursma [16] (2011)	The last 15 years of soccer in the Netherlands	Football-Data website [17]
Cao [18] (2012)	NBA League, 2005-2006 season until 2010-2011 season	Three valid basketball websites [8,11,19]

Table 2: The advantages and disadvantages of previous research

Researcher (Year)	*Advantages*	*Disadvantages*
Kahn [4] (2003)	Comparison of the designed system with previous systems, almost high prediction accuracy	Small dataset
McCabe and Travathan [6] (2008)	Participation in an international betting competition (Top-Tipper) in 2006-07 season	Low prediction accuracy
Zdravevski and Kulakov [7] (2010)	Collecting data from NBA league, designing the system as module	Low prediction accuracy, manual feature selection, absence of comparison with similar research
Trawinski [9] (2010)	Comparison between various fuzzy algorithms, calculating standard deviations of the algorithms, attending to the comprehensibility of the instructions	Small, local dataset, almost low prediction accuracy
Miljkovic et al. [3] (2010)	Up-to-date dataset, collecting data from NBA league	Low prediction accuracy
Davoodi and Khanteymoori [13] (2010)	Approximately high prediction accuracy	Small dataset
Ivankovic et al. [15] (2010)	Considering the effects of various shots on the results, evaluation of the future rival by determining the effects of each parameter on the winning result, approximately high prediction accuracy	Data collection from Serbia prevents public access to data and eliminates the chance for comparisons with similar research
Buursma [16] (2011)	Feature selection shed light on a number of features and parameters	Low prediction accuracy, absence of comparison with similar research
Cao [18] (2012)	Comprehensive and reliable dataset, automatic data collection and management, comparison with previous research	Low prediction accuracy, not updating the dataset, features depended on the last 10 games and games in the two first months cannot be used in training, absence of players' performance in prediction

Copyright (c) 2013 Advances in Computer Science: an International Journal. All Rights Reserved.

References

[1] Kent, M., “The Oxford Dictionary of Sports Science and Medicine”, Oxford university press, 1994.

[2] Schumaker, R., Solieman, O., Chen, H.” Sports data mining. Springer”, 2010.

[3] Miljkovic, D., Gajic, L., Kovacevic, A., Konjovic, Z., “The use of data mining for basketball matches outcomes prediction”, IEEE 8th International Symposium on intelligent and informatics, Subotica, Serbia, 2010, pp.309-312.

[4] Kahn, J., “Neural Network Prediction of NFL Football Games”, 2003, available on http:// homepages.cae.wisc.edu/~ece539/project/f03/kahn.pdf

[5] www.NFL.com, Retrieved Dec 2003.

[6] McCabe, A., Travathan, J., “Artificial Intelligence in Sports Prediction”, IEEE Computer Society Washington, DC, USA, 2008, pp. 1194-1197 .

[7] Zdravevski, E., Kulakov, A., “System for Prediction of the Winner in a Sports Game”, In: ICT Innovations 2009, Part 2, 2010, pp. 55–63.

[8] Basketball Reference, www.Basketball-reference.com

[9] Trawinski, K., “A fuzzy classification system for prediction of the results of the basketball games”, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 2010, pp.1-7.

[10] ACB league official site, www.ACB.com

[11] National Basketball Association (NBA) Official Website, www.NBA.com

[12] http://Sports.yahoo.com/NBA

[13] Davoodi, E., Khanteymoori, A.R., “Horse Racing Prediction Using Artificial Neural Networks”, Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing, 2010, pp.155-160.

[14] EQUIBASE company, www.equibase.com, Retrieved Jan 2010

[15] Ivankovic, Z., Rackovic, M., markoski, B., Radosav, D., Ivkovic, M., “Analysis of basketball games using neural networks”, 11th IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary, November 2010, pp.251-256.

[16] Buursma, D., “Predicting sports events from past results Towards effective betting on football matches”, Conference Paper, presented at 14th Twente Student Conference on IT, Twente, Holland, 21 January 2011.

[17] www.football-data.co.uk

[18] Cao, C., “Sports data mining technology used in basketball outcome prediction”, Master dissertation, Dublin Institute of technology, Ireland, 2012.

[19] Database Basketball, www.Databasebasketball.com, Retrieved July 2012.

[20] Purucker, M.C., ”Neural Network Quarterbacking", Potentials, IEEE, 15, 3,1996, pp. 9-15.

[21] Binesh, M., “Position of Data Mining in Knowledge Management”, Automobile Industry (In Persian), Vo. 36, 2009.

[22] O’Reilly, N., Knight, P., “Knowledge management best practices in national sport organizations”, International Journal of sport management and marketing, 2, 3, 2007, pp.264-280.

[23] Kantardzic, M., “Data mining – Concepts, models, methods, and algorithms”, Wiley-IEEE Press, 2003.

[24] Han, J., Kamber, M., Pie, J., “Data mining: Concepts and Techniques”, Morgan Kauffman Publishers, 2006.

[25] Landwehr, N., Hall, M., Frank, E., “Logistic model trees”, Machine Learning, 59, 1, 2005, pp.161–205.

[26] Heckerman, D., “Bayesian Networks for Data mining”, Data mining and Knowledge Discovery 1, Kluwer Academic Publishers, Manufactured in The Netherlands, 1997, pp.79-119.

[27] B. Korb, K., E. Nicholson, A., “Bayesian Artificial Intelligence”, Chapman and Hall/CRC, A CRC Press Company, 2003.

[28] Corporation, T.C., “Introduction to Data mining and Knowledge Discovery”, 1999.

[29] Kingsford, C., Salzberg, S.L., “What are decision trees?”, Nature Biotechnology, 2008, pp.1011-1013.

[30] Bart K., “Fuzzy Thinking”, K.N. Toosi University Publications(In Persian), 1998.

[31] Ahmadizadeh, S.S., Tabe, M., “An introduction to Fuzzy Models in User Specification or Locating with ArcSDM.”, (In Persian), 2010.

[32] Ye, N., “The handbook of data mining”, New Jersey: Lawrence Erlbaum Associates, 2003.

[33] Witten, I.H., Frank, E., “Data mining: Practical machine learning tools and techniques (2nd ed.)”, Morgan Kauffman Publishers, 2005.

[34] Palmer, A., Jimenez, R., Gervilla, E., “Data mining: Machine learning and statistical Technique”, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), 2011, pp.373-396.

[35] Hollinger, J., “Pro Basketball Prospectus”, Potomac Books, 2003.

Copyright (c) 2013 Advances in Computer Science: an International Journal. All Rights Reserved.