Basketball Game Outcome Prediction using Machine Learning

Harmandeep Kaur

Assistant Professor, Department of Computer Science and Engineering,

Chandigarh University, Mohali, Punjab, India

---***---

Abstract - Prediction of sports has always been an enthralling area for sports players, teams, fans and media and the growing number of gamblers. These days, large amount of effort and money has been spent by a number of companies to predict basketball game results, using machine learning. The presence of abundant data in sports and rapid growth of advanced technologies such as machine learning attracted a number of researchers for sports prediction. Support Vector Machine (SVM) is one of the most powerful techniques to efficiently handle classification problems. But they lack in rule generation. In the proposed work the Hybrid Fuzzy-SVM model (HFSVM) is developed to overcome the problem of SVM for rule generation that integrates the fuzzy approach with SVM technique to predict the basketball game results. The HFSVM model includes advantages of both SVM techniques (that is a unique strength of SVM) and fuzzy approach (which is rule generation ability. In the proposed work the developed HFSVM model is applied to the data of 1200 NBA games from 2015-2016 regular-season to predict basketball game outcome. The professional basketball game is becoming more and more popular due to its high-scoring and dynamic nature. At last, HFSVM model and SVM model are compared together and the outcomes show that HFSVM model achieves relatively satisfactory accuracy for prediction. Thus, HFSVM model can be used to obtain promising results when predicting outcome for basketball game.

Key Words: Membership function, Fuzzy Logic, Support Vector Machine, Basketball outcome prediction.

1. INTRODUCTION

Sports outcome prediction is a business that has popularly grown in recent years. Although, predicting sports game outcome is much difficult, but it has been given a close and large-scale attention for a long time. The result of a sports game is unknown till the end of the match. Due to such unpredictability of the outcome of sports games, the excitement of sports competition increases. Sports prediction is extremely important for coaches, sports fans, media and the growing number of gamblers. Due to rising demand for professional advice related to the sports event outcome, a variety of experts are involved in sports game prediction. Moreover, the presence of abundant data regarding the sports event's outcome makes it possible to perform significant research about the sports prediction. Researchers predict the outcome of sports events through a variety of simulation models, mathematical formulas or quantitative analysis. Two important sports prediction areas are: to get the factors affecting the game results and to know how profitable results can be obtained by changing these factors.

Basketball is one of the popular sports due to its highly dynamic nature. NBA, world's foremost level basketball league was established in 1946. At present, NBA includes total 30 teams out of which 29 teams are situated in the United States and one team is situated in Canada. The NBA has big following including experts anticipating results and abundant betting companies offering a large amount of money in gambling. Although basketball game has gained large popularity, it has received less attention in prediction areas.

For a long time, it has been the objective of many gamblers and researchers to precisely predict results of the sports based on the historical information. It has resulted in many sports-specific developments such as simulation models using statistical methods as well as machine learning models. Traditionally, researchers used simple statistical approaches to provide a team ranking list for predicting home team's probability to win the upcoming game by using statistics of past games played. But due to presence of ubiquitous data their accuracy is low. Zak et al. [1] combined offensive and defensive elements to rank individual team. It has been discovered that the field goal percentage, rebounding and free throw percentage give better results. Leake [2] applied least squares to obtain the ratings for college basketball game and professional and college football game. It was also found that the accurate prediction in college basketball, college football, and the professional basketball can be achieved by implementing least squares on a digital computer. Stefani [3] ranked the team by applying linear model to the score difference from each match and obtained the rankings by applying least squares. Whereas, Harville [4] applied a modified least square system to select college football and basketball teams for the postseason competition.

Spann and Skiera [5] made the comparison of the prediction accuracy using different methods which are prediction markets, betting odds and tipsters as well as evaluate the potential of tipsters and prediction markets to systematically generate the profits in a betting market. In terms of prediction accuracy, prediction markets and betting odds perform at an equal level. Both of them are better than tipsters. Forrest et al. [6] found that even in the case of financially pressured environment experts' subjective prediction is better than the prediction by statistical models. The experts' views represent, publish odds that are increasingly effective over a period of five years. Therefore, it has been demonstrated that the best source for games' probabilistic predictions are bookmaker odds.

Strumbelj and Vracar [7] applied possession-based Markov model to predict the progression of the NBA basketball game. The match was simulated using the model and the outcome prediction was produced. . It was found that Markov model approach was better than other statistical approaches while providing more insight onto the basketball. Vracar et al. [8] presented the methodology to generate simulation basketball match that is held between two different teams. Simulations involve a sequence of play-by-play in-game events at the team level. The results show that in a large part, the progression of a basketball game is a homogenous process except a few seconds at the beginning and at the end of each quarter. Modeling the non-homogenous part improves the prediction results and generates the simulations that better capture the dynamics of the basketball game's progression.

Statistical approaches are not an appropriate method for understanding issues in depth and for identifying the ways of solving the problem. They are also complex and time-consuming. Therefore, to overcome this problem machine learning techniques can be used that are more powerful and less time consuming. Cheng et al. [9] formalized NBA game outcome prediction problem as a classification problem and applied the Maximum Entropy principle to build an NBA Maximum Entropy model (NBAME). The results depict that the difficulty in the prediction of the NBA playoff outcomes is due to many unforeseeable factors such as the presence of the injured player, the relative strengths of the team, players' attitude and operations that determine the winner and loser by team's managers. Markoski et al. [10] developed a solution known as BBFBR (Basketball Board for Basketball referees) using neural network that takes the movement of the ball on the court as an input vector and the output vector of the neural network involves the movement coordinates of the referees. Ivankovic et al. [11] applied neural network on the data of the First B basketball league for men in Siberia and found that defensive rebound and two-point shots under the hoop are important elements in basketball. In defense, after opponent's shot, it is important to catch a ball and preventing from next offense while in the offense, to be precise under the hoop is considered important.

Support vector machine is the machine learning technique that can effectively handle classification problems [12-13]. But it still lacks in the ability of generating rules that are required for decision making [14]. The Hybrid Fuzzy-SVM model overcomes the advantage of SVM by generating rules using fuzzy approach. Although SVM technique is applied in a number of applications, but in many of these applications the input points may not be classified properly to one of the given class. The fuzzy membership functions when applied to SVM's each input point, then these input points tends to have a different significance to the decision surface learning. Therefore, fuzzy approach increases SVM performance by reducing the data inputs' noise effect and outliers that leads in the reduction of net error effect.

Balli and Korukoglu [15] developed a decision support framework to select the candidates that are eligible to become a basketball player by using Fuzzy Multi-Attribute Decision Making (MADM) algorithm. Trawinski

[16] presented a preliminary approach for creating a fuzzy model to predict the basketball game outcomes by using the KEEL (Knowledge Extraction based on Evolutionary Learning) system that selects, conduct and compare 10 fuzzy rule learning algorithms with a standard linear regression model. Wang [17] applied fuzzy regression on defence and offence in basketball game. It concludes that the significant factor to affect the game result was free throw percentage. In addition to this, free throw percentage is found to be most important among the attributes used, to affect the game and defence is more important factor than offense.

Pai et al. [18] developed a hybrid model by combining the SVM technique and the decision tree approach (HSVMDT) for the prediction of basketball game outcome and to help the coaches planning the strategies and players to enhance their performances. The model allows both forward and backward reasoning functions. The forward reasoning is used to predict basketball game results and the backward reasoning provides the advice to coaches to adjust the play strategy so as to win.

2. Proposed Framework

The proposed framework's flow diagram is shown in Fig. 1. The raw data are gathered from NBA websites. The raw data is then pre-processed for filling the missing values. In order to fill the missing values, firstly the data are segregated according to their data types and then by using caret algorithm, the missing values of the numeric data are imputed by using the caret algorithm. Before feature selection a complete data set is obtained by combining the segregated data. Selecting essential features plays very important role in the classification process as the features that do not affect the results are removed. This leads to the reduction of complexity at computational time and improvement of accuracy. In the proposed work, boruta algorithm is used to select important condition features. The features with their variable importance greater than the shadow variable importance are selected and the variables with lower variable importance are rejected. Number of iterations of boruta algorithm can be adjusted to select the tentative features. To reduce the variation in range of each feature the data is normalized and then classified. Classification plays a crucial role in predicting the results of the basketball game. The proposed work uses two different models, i.e. SVM model and HFSVM model are used to perform the classification. In SVM, firstly the processed data is loaded into the model. After partitioning the data into the training and testing data, the target feature and input features are set and the model is built on the training data by using a radial basis kernel function. The trained model is evaluated by using testing data. The model is evaluated on the basis of evaluation parameters: confusion matrix, accuracy, the time taken, sensitivity, specificity, precision and recall and results are represented in the form of plot. A four-fold cross validation is done to obtain average accuracy. At last, the results of the SVM model are saved. Therefore, this well-trained SVM model can be used to predict basketball game results that are useful for players and coaches for increasing performance in the game.

In HFSVM model the input data is fuzzified into the linguistic variables and their corresponding membership function is calculated that gives the extent to which the input value belongs to the fuzzy set. The rules created using fuzzy approach are evaluated and the aggregation approach is performed in which outputs of all the rules are unified. The rules are then defuzzified using a centroid approach to obtain crisp dataset. This dataset is then used as the input to build SVM model. In this way, the fuzzy approach is integrated with SVM technique. The remaining procedure to achieve the basketball game outcomes is same as followed by the SVM model. Finally, both models are compared using their prediction outcomes. The flowchart (Fig. 1.) explains the methods usend in the steps.

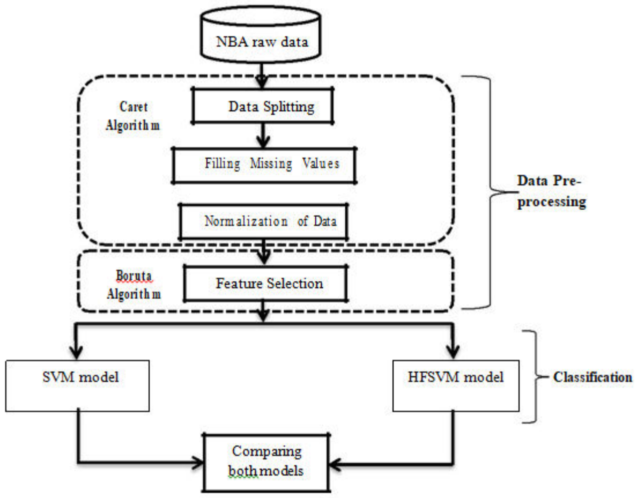

Fig.1. Flowchart of Proposed Framework

Fig -1: Flowchart of Proposed Framework

2.1 Data Pre-processing

The data pre-processing is a technique to remove the redundant values, impute missing values, removing correlated attributes and splitting as well as normalizing the data set. In this segment of the flow diagram, caret (classification and regression training) algorithm is applied for data pre-processing. The algorithm involves a set of functions so as to make the process of creating predictive models more effective and efficient. The algorithm provides the tools for performing the functions such as data splitting, finding missing values and normalizing the data. Data Splitting is done to select the random sample for the purpose of analysis. Balanced splits of the data are performed on the basis of result. After this, the algorithm performs analysis by selecting a random sample of data. Caret algorithm segregates the features of the data set according to their data types. Then the missing values are imputed in the data set of numeric data type. In this process, it estimates the features those are required for each operation and then applies them to a specific dataset. After filling the missing values, the caret algorithm normalizes the data set between zero and one. At last, the data set is combined and then it is passed to the feature selection process.

Feature selection methods are used to create models with different subsets of a dataset and determine the attributes that are necessary to build an accurate model. The important features are selected by using boruta algorithm. The boruta algorithm can be embedded with any classification model to obtain variable importance, but boruta algorithm uses random forest by default. Top down search is performed to get important features. The Z-Score variable importance of an input feature is compared to the Z-Score of shadow attribute that was generated by reordering original ones. The features with comparatively less variable importance than shadow attributes are rejected and the features with comparatively more variable importance are rejected. The execution of the algorithm in default light mode excludes unimportant features along with their random shadows. Whereas, the execution of the algorithm in force mode preserves all the features until the iterations are terminated. The boruta algorithm terminates under the two conditions: when last iteration is reached or when only confirmed features are left behind or when the tentative features are left without any decision. The number of iterations can be increased to avoid tentative features.

2.2 Classification Using Support Vector Machine (SVM)

SVM is a type of supervised learning algorithm that is used for classification, regression and outlier detection. SVM can classify both linear as well as nonlinear data. SVM model is used to avoid over-fitting. SVM can be more effective with a greater number of dimensions and the comparatively small number of samples. Since, SVM uses a subset of training points in the dataset (called support vectors)so it is memory efficient. Moreover, it is flexible as it has different kernel functions that can be used for the decision function.

The optimal separating hyperplane can be obtained on linearly separable data by using support vectors and margins. The hyperplane clearly separates different classes of the dataset. In case of nonlinear data, SVM converts the training data into a higher dimension by using nonlinear mapping. The optimal hyperplane separating different classes can be obtained by using the achieved higher dimension. Given the data as { , }, the is a set of the training tuples and is a set of the corresponding class labels for th set. Therefore, any point lying above the separating hyperplane satisfies the equation (1) and any point lying below the separating hyperplane satisfies the equation (2):

$$HP_1: \quad \omega.x_i + b_0 \geq 1, \quad if\ y_i = +1 \qquad (1)$$
$$HP_2: \quad \omega.x_i + b_0 \leq 1, \quad if\ y_i = -1, \quad for\ i = 1,2, \ldots \ldots, n \quad (2)$$

SVM used for binary class does not deal with slack variables. It depicts that input points belongs to class +1 if they fall on or above the hyperplane 1 and the input points belongs to the class −1 if they fall on or below the hyperplane 2 . From equation (1) and (2) we can obtain:

$$y_i(\omega.x_i + b) \geq 1, \qquad for\ i = 1,2, \ldots \ldots, n \quad (3)$$

By solving the quadratic problem, the margin between two classes can be maximized as given in equation (4):

$$Minimize, \quad \frac{1}{2}\ \omega.\omega \qquad (4)$$

The Lagrange multiplier is applied on equation (3) and (4) and Karush-Kuhn-Tucker (KKT) condition [19-20] is implemented to the solution to obtain the problem of optimization represented as:

$$Maximize, \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \quad (5)$$

The training tuples appear in the form of dot product when finding a linear SVM in the higher dimensional space. For obtaining linearly separable hyperplane from a nonlinearly separable problem the Kernel function is used. The kernel function maps the non-linearly separable data into feature space with higher dimension. The model uses the Gaussian Kernel function.

2.3 Classification Using Hybrid Fuzzy-SVM (HFSVM):

In HFSVM the fuzzy approach is used to generate the rules for decision making. The generated rules are used to build the SVM model and the rest of the process of classification and predicting the outcome is similar as SVM model. The fuzzy approach uses a Fuzzy Logic System (FLS), which is described as the nonlinear mapping of an input dataset to the scalar output data. The FLS involves four major components: fuzzifier, fuzzy inference engine, fuzzy rules and defuzzifier. These components form a general architecture of an FLS.

Fuzzifier. It converts the crisp input values to the linguistic variables using membership functions that are stored in the fuzzy knowledge base. Linguistic variables are fuzzy system's input or output variables that contain the value in the form of a word or sentence from a natural language, instead of having a numeric value. A membership function gives the measure of the degree to which the linguistic term belongs to a fuzzy set. In fuzzification the membership functions of fuzzy logic systems are used to map crisp input data to fuzzy linguistic terms.

Fuzzy Rules. In a Fuzzy logic system, a rule base is constructed to control the output variable. A fuzzy rule is the form of IF-THEN rule having a condition and a conclusion. The fuzzy set has the flexibility to model linguistic expressions. It expresses the degree to which a given element belongs to set. Fuzzy set operations perform fuzzy rule evaluations along with the combination of the results of the individual rules. The fuzzy set operations are OR fuzzy operation that is used to get the disjunction of the rule antecedents, AND fuzzy operation that is used to get the conjunction of the rule antecedents and NOT fuzzy that is used to get the complement of the rule antecedents.

Fuzzy Inference. When each rule is evaluated the final result is obtained by combining these rules which is known as inference. To merge the results of individual rules the accumulation methods are used that are maximum method, bounded sum method and normalized method. The inference step gives the result in the form of fuzzy value.

Defuzzifier. The result in the form of fuzzy value, obtained by fuzzy inference is converted into final crisp output. This process is performed by defuzzifier. Defuzification depends on the output variable's membership function. The algorithms used for defuzzification are Mean of Maximum (MOM) method, Bisector of Area (BOA) method and Centre of Gravity (COG) method.

3. Results and discussions

The proposed work involves the raw data that is collaborated from the websites such as "basketball-reference.com", "NBA.com". The raw data contains 1200 games from 2015-2016 regular seasons. There are 33 condition features and one decision feature in data. Feature selection is done using boruta algorithm using which 20 condition attributes are selected. The variable importance of the features against the shadow features is shown by boruta plot in Fig. 2. The green color in boruta plot represents the accepted features with high variable importance, whereas the tentative features are shown by the yellow color in the plot. It is shown that the features beyond FTr are accepted features are the accepted features and the rest of the features are tentative features.

After the selection of features, data is portioned into two parts, training data that includes 900 instances of the basketball data and testing data that includes 300 instances of the basketball data. The training data is used to train the model and then testing data is used to evaluate the model through the evaluation parameters. The evaluation parameters (used in the proposed work) and their values are given in Table 1. The true negative rate (also known as specificity) of SVM is 0.861 and that of HFSVM is 0.889. The true positive rate (also known as sensitivity or recall) of SVM is 0.79 and that of HFSVM is 0.924. It depicts that the HFSVM predicts more accurately the true 'LOSS' events than HFSVM. On the other side, SVM predicts more accurately the true 'WIN' events than SVM. But the accuracy depends on the net effect of sensitivity and specificity, which is more for HFSVM. Therefore, HFSVM is better in this case. The false positive rate of SVM is 0.139 and that of HFSVM is 0.111. SVM miss-classifies more 'WIN' events than HFSVM. The false negative rate of SVM is 0.21 and that of HFSVM is 0.076. Also, SVM miss-classifies more 'LOSS' events than HFSVM. The positive predictive rate of SVM (0.861) is more than that of HFSVM (0.901).

Table-1: Evaluation Parameters of SVM and HFSVM

	Abbreviation	SVM	HFSVM
TNR	True Negative Rate	0.861	0.889
TPR	True Positive Rate	0.79	0.924
FPR	False Positive Rate	0.139	0.111
FNR	False Negative Rate	0.21	0.076
PPV	Positive Predictive Rate	0.861	0.901

Table 2 shows the average Four-fold cross validation result of HFSVM and SVM models. The average testing accuracy of the HFSVM model (89.38%) is higher than the average testing accuracy (81.56%) of SVM model. The average computational time of HFSVM model is less than the average computational time of SVM model.

Table–2: Average Four-fold Cross Validation Results of HFSVM and SVM

	Average Accuracy	Average C/T(s)	Average No. of Support Vectors
HFSVM	8.938	1.42	573
SVM	81.56	2.05	743

The fourfold cross validation of SVM gives the average type 1 and type 2 prediction error rate 10.16% and 8%, respectively. However, the Four-fold cross validation of HFSVM gives the average type 1 and type 2 prediction error rate 4.6% and 6%,

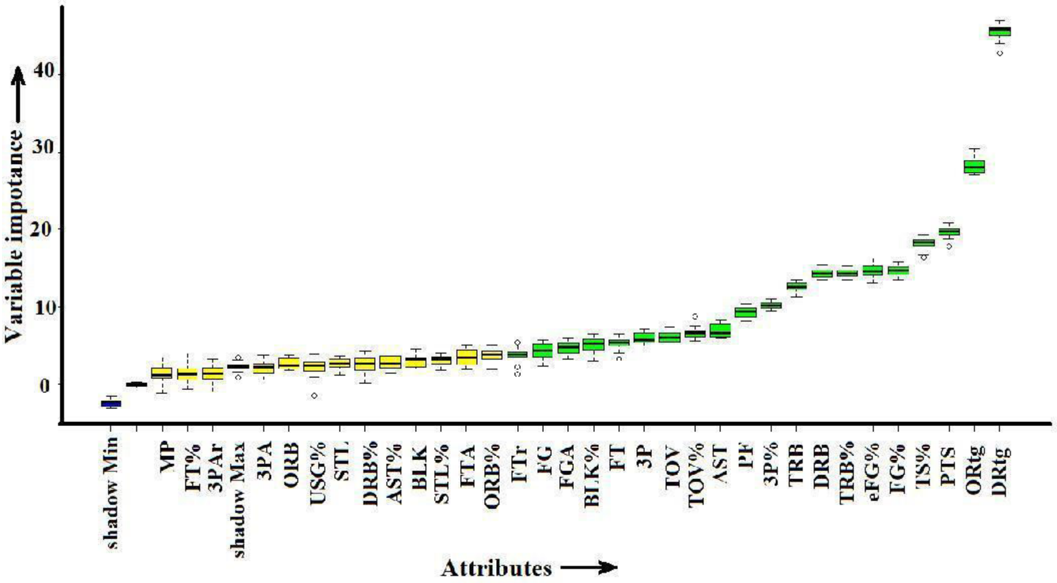

Fig -2 : Attribute versus Corresponding Variable Importance respectively. The total average error rate of HFSVM model is 10.6%, which is less than the SVM model's total average error rate that is 18.16%.Type 1 error represents the probability when the result by prediction model is "loss" but the true outcome is ''win". Type 2 error represents the probability when prediction model is "win" but the true result is "loss".Therefore, it can be concluded that the net error effect gets reduced on implementing fuzzy membership to each input point of the dataset because these input points make different contributions to the decision surface learning.

The testing accuracy in the previous paper [18] to predict the basketball game outcome was 85.25%. Thus, HFSVM attains quite adequate testing accuracy.

Artificial Intelligence and machine learning can be employed in different domains like drug discovery [21-22], fraud prediction [23-24], cancer prediction [25-26], etc. Authors in [27-29] describe the security and privacy aspects of the information especially the sensitive attributes like location and user identification present in the datasets used for empirical studies, while some good works discusses the same issue for discrete point dataset used for publishing the user data publicly [30-31].

4. Conclusion

Although SVM is a powerful classification tool but it lacks in rule generation ability. So, to overcome this problem HFSVM model is developed to predict the results of the games like NBA. It can be concluded that taking the basketball game's advanced features the accuracy of the model gets increased. On comparing both SVM and HFSVM model, it is discovered that HFSVM generates better results. The attributes that play very important role in predicting outcomes are defensive rating and offensive rating while the attributes minutes played and the field goal attempt rate has least effect on the results. The HFSVM model can result in higher average testing accuracy (89.38%) than that achieved by the SVM model (81.56%). The HFSVM model takes less computation time than the SVM model. Also, on comparing the HFSVM model with existing studies to predict the outcomes of basketball games, the HFSVM model provides quite adequate accuracy. Therefore, HFSVM model can be used as a promising substitute for predicting the basketball game results.

ACKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to Dr. Sushma Jain of Computer Science and Engineering, Thapar University, Patiala for giving me generous guidance, help and useful suggestions. Words fail me to express my thanks to my family for their selfless sacrifice, encouragement and heart full blessing that continue to enlighten my life and also thankful to my friends for their help and wishes. At last, I would express my special thanks to almighty god for blessing me with strength and wisdom to complete the research paper successfully.

REFERENCES

1. Zak TA, Huang CJ, Siegfried JJ (1979) Production Efficiency: The Case of Professional Basketball. The Journal of Business 52:379. doi: 10.1086/296053
2. Leake RJ (1976) A Method for Ranking Teams With an Application to College Football. in Management Science in Sports , eds. R. E. Machol, S.P. Ladany, and D.G. Morrison, Amsterdam: North-Holland 27-46.
3. Stern HS (1992) Who'S Number One? Rating Football Teams. In Proceedings of the Section on Statistics in Sports, American Statistical Association 1-6.
4. Harville DA (2003) The Selection or Seeding of College Basketball or Football Teams for Postseason Competition. Journal of the American Statistical Association 98:17–27. doi: 10.1198/016214503388619058

5. Spann M, Skiera B (2009) Sports forecasting: a comparison of the forecast accuracy of prediction markets,
   betting odds and tipsters. Journal of Forecasting 28:55–72. doi: 10.1002/for.1091
6. Forrest D, Goddard J, Simmons R (2005) Odds-setters as forecasters: The case of English football. International Journal of Forecasting 21:551–564. doi: 10.1016/j.ijforecast.2005.03.003
7. Jian W, Zhi-Hua H, Zhi-Yong Z (2014) Clustering Analysis of Sports Performance Based on Ant Colony Algorithm. 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications. doi: 10.1109/isdea.2014.71
8. Vračar P, Štrumbelj E, Kononenko I (2016) Modeling basketball play-by-play data. Expert Systems with Applications 44:58–66. doi: 10.1016/j.eswa.2015.09.004
9. Cheng G, Zhang Z, Kyebambe MN, Kimbugwe N (2016) Predicting the Outcome of NBA Playoffs Based on Maximum Entropy Principle. doi: 10.20944/preprints201609.0103.v1
10. Markoski B, Pecev P, Ratgeber L, Ivković M, Ivanković Z (2011) A new approach to decision making in basketball-BBFBR program. ActaPolytechnicaHungarica. 8(6):111-30.
11. Wang, Kuan-Chieh, Richard Z (2016) Classifying NBA offensive plays using neural networks. Proc. MIT SLOAN Sports Analytics Conf.
12. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20:273–297. doi: 10.1007/bf00994018
13. Vapnik, V (1995) The nature of statistical learning theory Springer New York Google Scholar.
14. Diederich J (2008) Rule Extraction from Support Vector Machines: An Introduction. Rule Extraction from
    Support Vector Machines Studies in Computational Intelligence 3–31. doi: 10.1007/978-3-540-75390-2_1
15. Ballı S, Korukoğlu S (2012) Development of a fuzzy decision support framework for complex multi-attribute decision problems: A case study for the selection of skilful basketball players. Expert Systems 31:56–69. doi: 10.1111/exsy.12002
16. Trawinski K (2010) A fuzzy classification system for prediction of the results of the basketball games. International Conference on Fuzzy Systems. doi: 10.1109/fuzzy.2010.5584399
17. Wang JN (2000) The fuzzy regression analysis application of offensive and defensive techniques in the basketball game. Natl SciCouncRepub China Part C HumanitSocSci 10:287-298.
18. Pai P-F, Changliao L-H, Lin K-P (2016) Analyzing basketball games by a support vector machines with decision tree model. Neural Computing and Applications. doi: 10.1007/s00521-016-2321-9