Machine Learning Approaches to Predict Basketball Game Outcome

Harmandeep Kaur
CSED
Thapar University
Patiala-147004, India
hkaur2_me15@thapar.edu

Sushma Jain
Assistant Professor, CSED
Thapar University
Patiala-147004, India
sjain@thapar.edu

Abstract: **-Sports prediction has always been a spellbinding research area for sports fans, teams, team managers, team players and a growing number of gamblers.Nowadays, companies are spending more effort in machine learning to predict the sport outcomes. Support Vector Machines (SVMs) are powerful techniques that handle classification problems effectively and efficiently. However, SVM models lack in rule generation. So, this examination leads towards the development of Hybrid Fuzzy-SVM model (HFSVM) by integrating fuzzy approach and SVM technique for prediction ofthe basketball game outcomes that help the coaches, teams and players to enhance their performance. The HFSVM model combines the advantage of both SVM technique and fuzzy approach, which is a unique strength of SVM and rule generation ability of fuzzy approach using fuzzy membership functions. The HFSVM model is compared with SVM model and the empirical results showed that the HFSVM model not only provides better results than SVM model but also provides relatively satisfactory prediction accuracy. Therefore, promising results can be obtained using HFSVM model when analyzing the outcomes of basketball competitions.**

Keywords—Fuzzy Logic, Support Vector Machines, Machine Learning, Prediction

I. INTRODUCTION

A large amount of effort is spent to forecast the sporting event outcome. Two important strands of sports forecasting are: to obtain the factors that affect the game result and to learn how these factors can be changed so that profitable results can be obtained. The rapid advancement of high performance computing devices and the presence of abundant data in sports attract more academic attention towards quantitative study of professional sports.

The National Basketball Association(NBA), the leading men's professional basketball league in the world, established in 1946. The NBA has many followers around the world, with contenders predicting the result, in addition to numerous betting companies that are offering a large amount of money to predictors [1]. Professional basketball games are one of the most studied areas of many quantitative researches due to its popularity as well as dynamic and high scoring nature.

In previous decades, researchersapplied simple statistical principles that only combined technical features of past games played, to generate the team ranking list used in forecasting the probability of home teams in winning the upcoming game [2]. Their accuracy is low as data became more ubiquitous. It has been suggested that bookmaker odds are the best source of probabilistic forecasting of sports games [3], [4]. Strumbelj and Vracar assumed Markov property and used logistic regression to model state transitions [5], [6]. Although the model deals with the non-homogeneity of progression of a basketball game, but it was found that the summary statistics that were used, do not explain the team characteristics well.

Zak, Huang and Siegfried ranked individual team by combining offensive and defensive elements [2]. In [7], [8]it has been reported that one well established approach to rank the team is to apply the linear model to the score difference from each match and least squares to obtain the rankings. Wang concluded that the essential factor to affect the game outcome is free throw percentage. It was also found that defense is more essential than offense [9].

Machine learning algorithms overcome the disadvantages of statistical models by creating data driven predictions or decisions using a model from sample input. It has strong bonds to mathematical optimization. Plenty of companies are investing heavily in machine learning to predict the sports results. But this technique has not been extensively used in the basketball game analysis. It is found that Support Vector Machines are one of the most powerful classification techniques of machine learning [10], [11]. Although the SVM technique gives effective result for classification problem, but this approach has main disadvantage that it is incapable of yielding rules for decision making [12].

This disadvantage can be overcome by using Hybrid Fuzzy-SVM (HFSVM) model for rule generation. Nowadays, a large number of applications are using the SVM Technique. However, in many applications, some of the input points may not be exactly appointed to one of the given classes. So, they need to be assigned toone of the given class, for SVM to perform classification of these points more correctly. In this paper, a fuzzy membership function is applied to each of the input points of SVM. These input points make different contributions to the decision surface learning. Thus, it enhances the SVM to reduce the effect of noises and outliers in data inputs which directly reduces the net error effect.An overview of modeling techniques using fuzzy logic, various agriculturalandbiological systems and an example to show

978-15090-6403-8/17/$31.00 © 2017 IEEE

step-wise construction of a fuzzy logic model is provided by the research [13]. Hartati proposed a fuzzy based decision support system (DSS) to evaluate the suitability of land and to select the type of crop to be grown on it using climate, soil and land preparation as the parameters [14].

The rest of the research is arranged as follows. Section 2 introduces the proposed framework. In Section 3, experiment and results are discussed and the conclusion is given in Section 4.

II. PROPOSED FRAMEWORK

The flow diagram of proposed frame work is shown in Fig.1. The first step is the collection of raw datafrom NBA websites. The second step is data pre-processing. In this step, data segregation is performed according to the data types and after that missing values are imputed by using Caret algorithm. The third step is thefeature selection process. Feature selection is a crucial step before the classification process so as the features that are unimportantare deleted from the original dataset. This reduces computational complexity andincreases accuracy. In the proposed model, Boruta algorithm is employed to select the essential condition attributes.The attributes with their variable importance higher than shadow variable importance are selected. The third part of the work is to classify the data. Here, classification is done by using two different ways that is using an SVM model and HFSVM model. In SVM, the processed data is partitioned into a training dataset and testing dataset. The training data set is used to model SVM with essential attributes and classification performance of the trained SVM model is evaluated by employing testing dataset on a model that is trained. A fivefold cross validation is carried out to obtain averageaccuracy. Thus, the well trained SVM model can be used for forecasting the competition result that can be used bycoachesorplayerstoincrease

978-15090-6403-8/17/$31.00 © 2017 IEEE

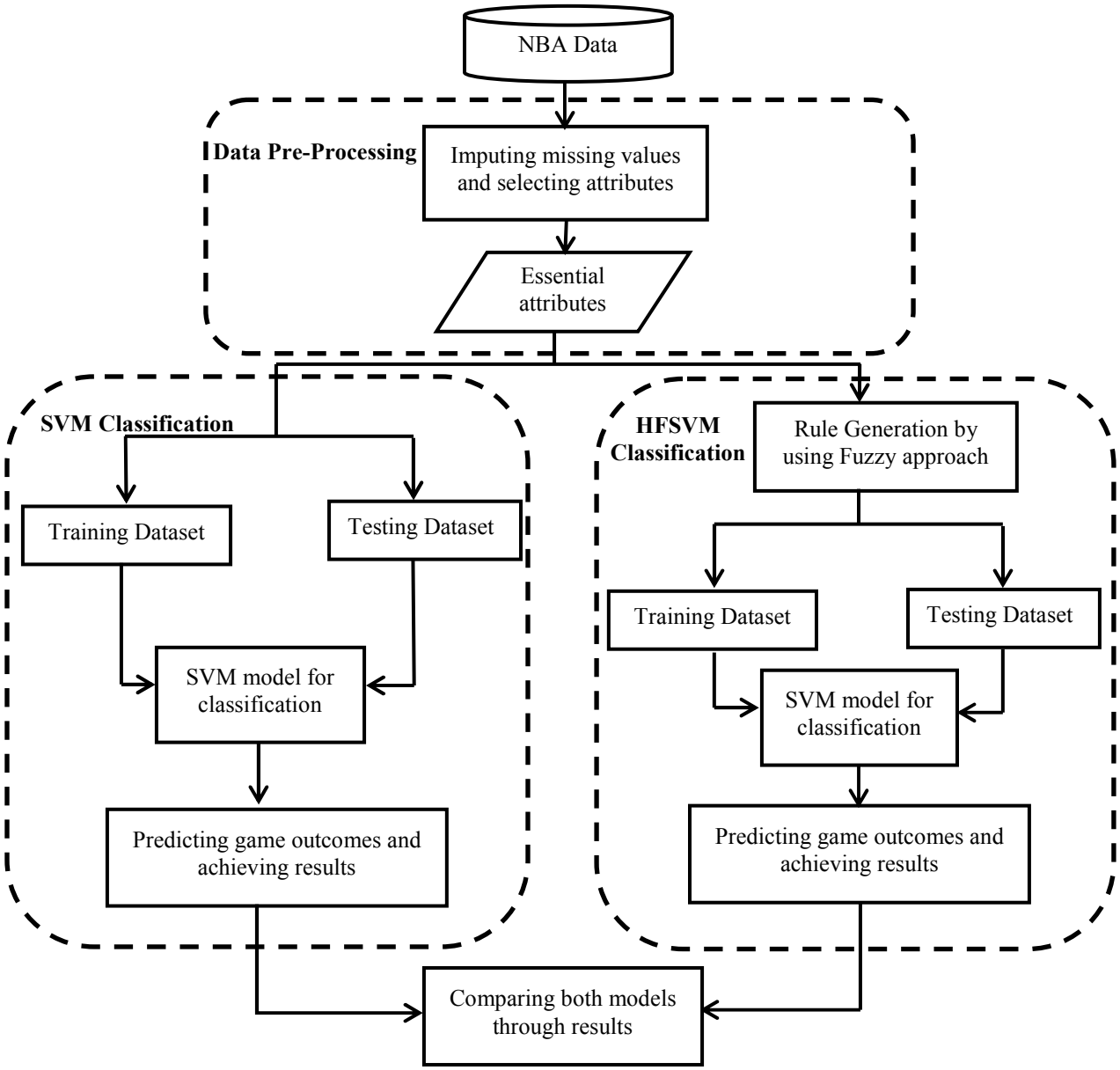

Fig. 1. Schematic diagram of Proposed Framework

performance in the game.

In HFSVM model rules are generated using fuzzy approach. Generated fuzzy rules are then evaluated and defuzzifiedto obtain crispdataset. This dataset is then used as the input to buildthe SVM model. This is how the fuzzy approach is integrated with SVM technique. The rest of the process to obtain the basketball game results is same as followed bythe SVM model. At last, both models are compared through their prediction results.

The methods used in the steps of the flowchart, given in Fig. 1, are explained as:

A. Data Pre-processing

This segment from flowchart utilizesCaret algorithm (short for classification and regression training) which includes a set of functions that attempts to make the process for generating predictive models more efficient and effective. This algorithm contains tools for functionalities such as data splitting, feature selection, pre-processing and variable importance estimation. It creates balanced splits of data based on the outcome. After splitting data, it selects the random sample of dataset for analysis and then segregates the attributes in dataset according to their data types. For numeric attributes Caretperforms pre-processing to impute missing values. In this process it estimates the parameters that are required for each operation and applies them to specific dataset. It imputes the dataset based only on information about training dataset.

B. Attribute Selection

In this segment, Boruta algorithm is used to selectrelevant features. It can work with any classification

978-15090-6403-8/17/$31.00 © 2017 IEEE

method that output variable important measures, but by default Boruta uses Random Forest. In each iteration,Boruta compares Z-Score of a condition attribute with Z-Score of shadow attribute that are created by rearranging original ones. Attributes having significant worst importance than shadow attributes are consecutively eliminated and attributes with significantly better importance are accepted. Algorithm running in default light mode drops unimportant attributes along with their random shadows. On the other hand, in the force mode all shadow attributes are preserved until the termination of iterations. The algorithm stops on two conditions: when only confirmed attributes are left or when last iteration is reached. In case of second condition the attributes may be left without a decision, called tentative attributes. To avoid that, the number of iterations can be extended.

C. Classification

The classification is proposed in two different ways:

1) Support Vector Machine (SVM): Support vector machine is a technique that is used to classify both linear and nonlinear data. SVMs are capable of modeling complex nonlinear boundaries and are highly accurate. They can be used for classification as well as numeric prediction and avoid over fitting. SVM can perform linearly inseparable classification with global optimal.

When using linear separable data they are used to get optimal separating hyperplane with the help of support vectors and margins. This hyperplane depicts the clear separation of different classes in the dataset. For nonlinear data SVM uses nonlinear mapping to convert existing training dataset into higher dimension. By using this new dimension, SVM searches for linearly optimal hyperplane separating the classes of the dataset. So, in case of any clutter function SVM finds some feature space. Given the dataset D as$\{x_i, y_i\}$, $i = 1,2,\dots\dots,n$, $y_i \in \{+1, -1\}$where $x_i$is the set training tuples corresponding to their class labels y_i. SVM discovers the hyperplane with the largest margin called Maximum Marginal Hyperplane (MMH). This margin gives the maximum separation between the classes. There exists a pair (w, b) such that w is a weight vector and b is a scalar bias that satisfies the condition [10], [11]:

$$H_1: \quad w.x_i + b \geq 1 - \varepsilon_i, \quad if\ y_i = +1 \tag{1}$$

$$H_2: \quad w.x_i + b \leq 1 - \varepsilon_i, \quad if\ y_i = -1, \tag{2}$$

$$for\ i = 1,2,\dots\dots,n$$

Whereε_i, is a negligence variable that allows error in the training set and also makes a soft marginal hyperplane for classification.This slack variable is zero for two class separation. This indicates that tuple falling on or above hyperplane H_1 belongs to class +1 and tuple falling on or below H_2 belong to class -1.Combining both (1) and (2) we get:

$$y_i(w.x_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 1, \tag{3}$$
$$for\ i = 1,2,\dots\dots,n$$

Tuples falling on H_1 and H_2 hyperplanes are called support vectors. These are tuples that are the most difficult to classify, but they provide more information regarding classification. The margin separating two classes can be maximized by solving the following quadratic problem.

$$Minimize \quad \frac{1}{2}\, w.w + C\sum_{i=1}^{n} \varepsilon_i \tag{4}$$

Subjected to the (3), where C is the constant used for controlling the scale of the margin and classification error. Equation (3) and (4) can be solved by using Lagrange multiplierα and then implementingKarush-Kuhn-Tucker (KKT) condition [15], [16] to the solution, the optimization problem can be written as:

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

$$\text{Subjected to,} \quad 0 \leq \alpha_i \leq C, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{6}$$

$$for\ i = 1,2,\dots\dots,n$$

The kernel function $k(x_i, x_j) = \phi(x_i) \times \phi(x_j)$is used for converting non-linearly separable problem to the linearly separable problem by mapping non-linearly distinguishable data into higher dimensional feature space.After finding a solution, the distinguishing function can be given by:

$$d(x_i) = sign\left[\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b\right], \tag{7}$$
$$i = 1,2,\dots\dots,n$$

The kernel function usedin (7) is aGaussian function given by the following (8):

$$k(x_i, x_j) = e^{\left[-\frac{(x_i - x_j)^2}{2\sigma^2}\right]} \tag{8}$$

2) Hybrid Fuzzy-SVM (HFSVM): Fuzzy logic is a system that is a form of multi-valued logic which includes any real number between 0 and 1 as the truth value of variables. In fuzzy logic, linguistic variables are used to facilitate the rules and facts expression. Fuzzification operation maps mathematical input values into membership functions. Membership functions allow quantifying linguistic variables with degree of membership. Finally, fuzzy rules are generated as well as evaluated and defuzzification is used to map a fuzzy output into a crisp output. The output after defuzzification is used as the input to build SVM model. In this way the fuzzy approach is integrated with SVM technique. The rest of the process to attain the basketball game outcomes is same as followed in the SVM model discussed earlier.

III. Experiment and Results Discussion

In proposed work, the dataset used is collected from some websites such as "NBA.com", "basketball-reference.com"which provide very valuable and informative data.

The NBA played in Canada and USA is divided into two conferences that is Western and Eastern conference. Each conference has 3 divisions and each division has 5 Teams each. In NBA each team plays 82 games in the regular season. 30 games are scheduled against non-conference opponents. There are 36 games that scheduled within the conference, but out of division, that is team plays 4 games with each of 6 teams within the conference and 3 games with each of 4 other teams within the conference. And also 16 games are scheduled within the division. A team plays with its 4 division arrivals 4 times each.

In this study, Data with total 800 games is collected from 2015-2016regular seasons. This data contains one decision attribute and 33 are condition attributes. The attributes employed in this study are depicted in Table I along with their corresponding description. First nineteen attributes in this table are basketball game's fundamental attributes and rests of the attributes except a decision attribute, are advanced attributes of the game. The attributes from X1 to X33 are condition attributes and Y is a decision attribute.

TABLE I.ATTRIBUTES OF NBA USED IN PROPOSED WORK

Attributes	**Abbreviation**	**Description**
X1	MP	Minutes Played
X2	FG	Field Goal
X3	FGA	Field Goal Attempts
X4	FG%	Field Goal Percentage
X5	3P	3-Point Field Goal
X6	3PA	3-Point Field Goal Attempts
X7	3PA%	3-Point Field Goal Percentage
X8	FT	Free Throw
X9	FTA	Free Throw Attempts
X10	FT%	Free Throw Percentage
X11	ORB	Offensive Rebound
X12	DRB	Defensive Rebound
X13	TRB	Total Rebound
X14	AST	Assists
X15	STL	Steals
X16	BLK	Blocks
X17	TOV	Turnover
X18	PF	Personal Fouls
X19	PTS	Points
X20	TS%	True Shooting Percentage
X21	eFG%	Effective Field Gaol %age
X22	3PAr	3-Point Attempt Rate
X23	FTr	Free Throw Attempt Rate
X24	ORB%	Offensive Rebound %age
X25	DRB%	Defensive Rebound %age
X26	TRB%	Total Rebound Percentage
X27	AST%	Assist Percentage
X28	STL%	Steal Percentage
X29	BLK%	Block Percentage
X30	TOV%	Turnover Percentage
X31	USG%	Usage Percentage
X32	ORtg	Offensive Rating
X33	DRtg	Defensive Rating
Y	W/L	Win/Loss

The Boruta algorithm is used for feature selection. It provides the result in the form of a plot with the importance of attributes against shadow attributes as shown in Fig. 2.The attributes with high variable importance are shown in green color, the attributes in yellow color are

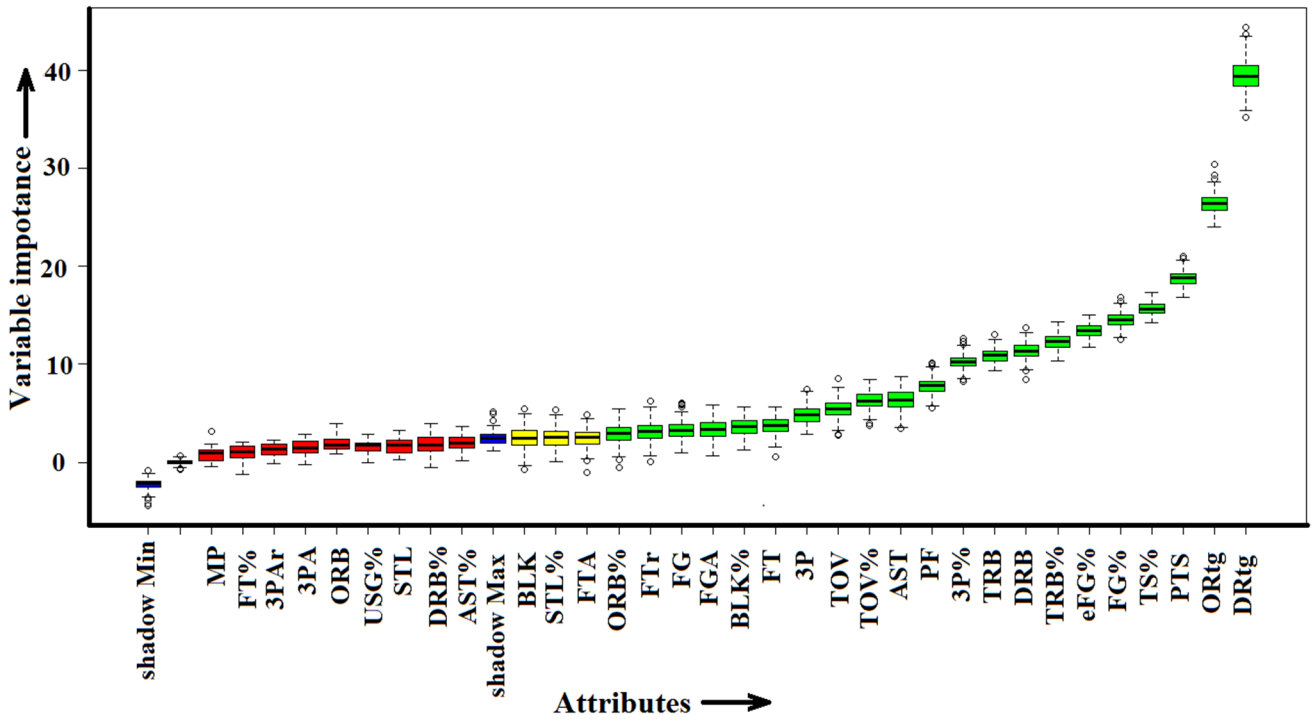

Fig. 2. Attributes v/s Corresponding variable importance plot by Boruta

tentative attributes and attributes in red color are rejected attributes and are of low variable importance. This means that the attributes beyond ORB% are selected attributes. BLK, FTA and STL% are the tentative attributes and remaining attributes are rejected. The Boruta algorithm selects 21 condition attributes. After selection ofattributes, the partitioning of data is done in two parts: the training dataset contains 640 games and testing dataset contains 160 games of the regular season. Table II and III represent the result for each of the classifications that are SVM and HFSVM respectively. The measuring parameters are: accuracy, computational time (in seconds)and number of support vectors for both table II and table III. The highest testing accuracy (87.82%) for the SVM model is given in Table II by Cross Validation CV2. Similarly, the highest testing accuracy (89.26%) for the HFSVM model is given in Table III by Cross Validation CV3.

TABLE II. EXPERIMENT RESULTS OF TESTING ACCURACY WITH SVM MODEL

	Accuracy	C/T(s)	Number of Support Vectors
CV1	85.71	1.84	540
CV2	87.82	1.79	545
CV3	85.09	1.75	538
CV4	85.71	1.74	551
CV5	86.34	2.04	559

TABLE III. EXPERIMENT RESULTS OF TESTING ACCURACY WITH HFSVM MODEL

	Accuracy	C/T(s)	Number of Support Vectors
CV1	87.60	0.93	390
CV2	88.43	0.94	401
CV3	89.26	1.09	381
CV4	87.26	1.22	392
CV5	88.43	102	393

Table IVoutlines average testing accuracy, average computation time and average number of support vectors of both SVM and HFSVM models. In the table, it is deducted that HFSVM model can achieve higher average testing accuracy (88.26%) than can the SVM model (86.21%). The computation time of the HFSVM model is shorter than the SVM model. Therefore, we can conclude that the net error effect is reduced when fuzzy membership is implemented to each input point of dataset as these input points make differentcontributions to the decision surface learning.

TABLE IV. AVERAGE FIVEFOLD CROSS VALIDATION RESULTS OF SVM AND HFSVM

	Average Accuracy	Average C/T(s)	Average No. of Support Vectors
HFSVM	88.26	1.04	391
SVM	86.21	1.83	547

978-15090-6403-8/17/$31.00 © 2017 IEEE

In the previous paper [17], the testing accuracyfor predicting outcomes of basketball games was (85.25%). Thus, the testing accuracy attained by the HFSVM model is quite adequate. In addition to this, the average type 1 and type 2 prediction error rates of fivefold cross-validation with SVM are 6.21% and 7.56% correspondingly. SVM model's total average error rate is 13.77%. The average type 1 and type 2 of fivefold cross-validation with HFSVMare 6.44% and 5.28% correspondingly. HFSVM model's total average error rate is 11.72%, which is less than the total average error rate of SVM. Also, in HFSVM the type 2 error is smaller than type one error which is considered better in predicting outcome. Type 1 error depicts the probability when the true outcome is ''win" but the result by prediction model is "loss". Type 2 error depicts the probability when the true result is "loss", but the result but prediction model is "win".

IV. Conclusions

Analyzing the basketball game is an interesting research area for the researchers because of the arousing curiosity of fans and social media for the prediction of outcomes of the basketball competitions. This is also helpful for the teams and players to enhance their performance. But predicting the outcome of the basketball game is a challenging task. It is found that taking advanced attributes of NBA game results in increasing the accuracy of the model. So, this leads towards the development of the HFSVM model for predicting the outcome in the NBA. Comparing SVM model with HFSVM model, it is found that HFSVM provides better results. Also, on comparing the HFSVM model with previous studies in predicting the basketball games'outcomes by SVM, the accuracy achieved by HFSVM model is quite adequate. Therefore, HFSVM model can be used as a promising alternative for predicting the basketball game outcomes.The HFSVM model is only capable to predictingthe win and loss outcomes of basketball games. Further extending HFSVM model the scores of win and loss can be investigated. This model can also be further applied to other sports such as soccer, baseball and golf, in order to check the feasibility of HFSVM model.

References

[1] D. Andrews, "The (Trans) National Basketball Association: American commodity-sign culture and global-local conjuncturalism." *Articulating the Global and the Local. Boulder, CO: Westview*, pp. 72-101, 1997.

[2] T.A. Zak, C.J. Huang, and J.J. Siegfried, "Production efficiency: the case of professional basketball." *Journal of Business*, pp. 379-392, 1979.

[3] C. Song, B.L. Boulier, and H.O. Stekler, "The comparative accuracy of judgmental and model forecasts of American football games." *International Journal of Forecasting*, vol. 23(3), pp. 405-413, 2007.

[4] M. Spann and B. Skiera, "Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters." *Journal of Forecasting*,vol. 28(1), pp. 55-72, 2009.

[5] E. Štrumbelj and P. Vračar, "Simulating a basketball match with a homogeneous Markov model and forecasting the outcome." *International Journal of Forecasting*, vol. 28(2), pp. 532-542, 2012.

[6] P. Vračar, E. Štrumbelj, and I. Kononenko, "Modeling basketball play-by-play data." *Expert Systems with Applications*, vol. 44, pp. 58-66, 2016.

[7] R.T. Stefani, "Football and basketball predictions using least squares." *IEEE Transactions on systems, man, and cybernetics*, vol. 7, pp. 117-121, 1977.

[8] D.A. Harvilleand M.H. Smith, "The Home-Court Advantage: How Large is it, and does it vary from Team to Team?." *The American Statistician*, vol. *48*(1), pp.22-28, 1994.

[9] W. ChingNan, "The application of fuzzy regression on offence/defense in basketball games." *Proceedings of the National Science Council, Republic of China. Part C, Humanities and Social Sciences*, vol. *10*(3), pp.287-298, 2000.

[10] C. CortesandV. Vapnik, "Support-vector networks." *Machine Learning*, vol. *20*(3), pp. 273-297, 1995.

[11] Vapnik,*The nature of statistical learning theory*. Springer science & business media, 2013.

[12] N., BarakatandA.P. Bradley,"Rule extraction from support vector machines: a review." *Neurocomputing*, vol. *74*(1), pp. 178-190, 2010.

[13] B. Center and B.P. Verma, "Fuzzy logic for biological and agricultural systems." *Artificial Intelligence Review*, vol. *12*(1-3), pp. 213-225, 1998.

[14] S. Hartati and I.S. Sitanggang,"A fuzzy based decision support system for evaluating land suitability and selecting crops."*Journal of Computer Science,*vol. 6(4), pp. 417-424, 2010.

[15] W. Karush, "Minima of functions of several variables with inequalities as side conditions." *Traces and Emergence of Nonlinear Programming*, pp 217-245, 2014.

[16] H.W. Kuhn andA.W. Tucker, "Nonlinear programming."*Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press,* pp. 481-492, 1951.

[17] P.F. Pai, L.H. ChangLiao, andK.P. Lin, "Analyzing basketball games by a support vector machines with decision tree model." *Neural Computing and Applications*, pp. 1-9, 2016.