STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **8**, June 2020, pp 602–609.
Published online in International Academic Press (www.IAPress.org)

Heuristics for Winner Prediction in International Cricket Matches

V Sivaramaraju Vetukuri [1,*], Nilambar Sethi [2], R. Rajender [3]

[1] *Research Scholar, Department of Comp. Sc. & Engg, Biju Patnaik University of Technology Rourkela, Odisha-769004, INDIA*
[2] *Dept. of Comp. Sc. & Engg, GIET, Gunupur Odisha-765022, INDIA*
[3] *Dept. of Comp. Sc. & Engg, Lendi Institute of Engineering & Technology, Visakhapatnam-531173, A.P, INDIA*

Abstract Cricket is popularly known as the game of gentlemen. The game of cricket has been introduced to the World by England. Since the introduction till date, it has become the second most ever popular game. In this context, few a data mining and analytical techniques have been proposed for the same. In this work, two different scenario have been considered for the prediction of winning team based on several parameters. These scenario are taken for two different standard formats for the game namely, one day international (ODI) cricket and twenty-twenty cricket (T-20). The prediction approaches differ from each other based on the types of parameters considered and the corresponding functional strategies. The strategies proposed here adopts two different approaches. One approach is for the winner prediction for one-day matches and the other is for predicting the winner for a T-20 match. The approaches have been proposed separately for both the versions of the game pertaining to the intra-variability in the strategies adopted by a team and individuals for each. The proposed strategies for each of the two scenarios have been individually evaluated against existing benchmark works, and for each of the cases the duo of approaches have outperformed the rest in terms of the prediction accuracy. The novel heuristics proposed herewith reflects efficiency and accuracy with respect to prediction of cricket data.

Keywords Cricket, T-20, ODI cricket, Data mining, Winner prediction.

AMS 2010 subject classifications 94A08, 62-07

DOI: 10.19139/soic-2310-5070-648

1. Introduction

Cricket is the game of gentlemen introduced to the world by the Britishers. It has gone through several modifications since its introduction, and now it exhibits in three different formats. These are test cricket (a five days format), ODI (One day international), and T-20 (An aggressive format comprising of a total of 40 overs). Among these, the ODI and T-20 formats have gained a lot of popularity and thus attracted the attentions of billions of audience and numerous business firms. Many of the countries with high density of populations are involved in the game. Few of the names are India, Pakistan, England, Australia, and Srilanka. The prediction of winning team for a forthcoming match is too much crucial for three most important aspects as mentioned below:-

- Essential for a country to strengthen the team,
- Essential for a business firm to invest,
- Essential for the team itself to adopt new strategies.

Data mining and machine learning have been adopted as essential tools for several aspects like classification, recognition, data analysis, and prediction in the field of computing. Numerous techniques have been proposed so far in this context ([1], [2], [3], [4], [5], [6]) irrespective of the research areas (medical, weather, big data,

[*] Correspondence to: V Sivaramaraju Vetukuri (Email: sivaramaraju.vetukuri@gmail.com). Research Scholar, Department of Computer Science and Engineering Biju Patnaik University of Technology Rourkela, Odisha-769004, INDIA

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2020 International Academic Press

education, business, banking, etc.). Recent study reveals various heuristics with utmost efficiency being presented in almost every research domain. However, there exist certain research domains where there exist multiple scopes for implementing these data mining approaches and analyze the outcomes. The game of cricket is one among such domain. Several data mining techniques have been proposed so far for different aspects involved in the match of a cricket. However, it is still a challenging task for proposing an efficient prediction heuristic. In [7], a statistics based model has been presented for the suitable selection of players for a particular team. Most essentially the past performance of the players have been considered as the basis. The batting, bowling and overall statistics pertaining to individual players have been considered for the work. The last five match performance only have been considered in this work. On a Hadoop setup, an accuracy of 91% have been reported for the work only for the Indian players. Another method of player selection has been proposed in [8]. For the purpose, neural network approach has been utilized. They have considered the historical match statistics during the year 1985-2006. Progressive training and testing has been done on four different sets of data. In this work also only the recent player performance during world cup have been taken into account. However, debut-ant players characteristics can not be analysis as the method ignores the same. In citea3, analysis has been made on the powerplay characteristics during a cricket match irrespective of the match format. The analysis has taken into account the difference between the score if there is a powerplay during the match and if there is no powerplay during that match. The themes around this work includes various powerplay formats, benefits of the powerplay for the batting team, benefits of the powerplay for the bowling team, and nature of the match without powerplay. However, powerplay strategies vary between the ODI and T-20 format of a match. COnsidering no powerplay is also an hypothetical situation that may not be fitting suitable for every model of analysis. A cricket outcome predictor has been presented in [10]. ODI outcome is being predicted using this method. Several feature considered for the work are nature of the match (day or day/night match), index of the innings (1st/ 2nd), and fitness of the teams. Classifiers used for the work are Naive bayes, support vector machine (SVM), and Random forest (RF). Combining these three classifiers outcomes, a tool has been developed namely COP (cricket outcome predictor). However, quantification of certain features considered here is a tedious task. Also, this work does not predict the outcome of a T-20 format match. In [11], a forecasting model has been proposed for runtime prediction of the outcome of an ODI cricket match. Logistic regression has been used as the basic tool. The work does the prediction with minimal number of features because of the use of a cross-validation technique, they have eradicated features with less importance. A study has been made in [12] for determining the importance and usefulness of business betting for the match of cricket. They have suggested a profit of 20% is achievable if netting is done as per the fall of of wickets during the match. The Monte-Carlo estimation has been used for the purpose.

In [13], predictive tool has been presented for the test cricket match format. A test cricket match is played between two teams for a duration of five days with each day being played for around 90 overs. They have used a probabilistic approach for the prediction of the final outcome. Twelve different precondition parameters have been considered in this work. It also uses the logistic regression as it's basis tool. In [14], online social information have been used for making a prediction of the top ranking players and teams in cricket. The future trend of a match is generated based on the data trending on social media. In [15], two different themes have been merged into a model for predicting of outcome of a ODI match. Based on various parameters pertaining to the first and second innings of a match (50 overs each), the outcomes are generated at runtime. Linear regression and Naive Bayes have been utilized for the said purpose. A mild rate of accuracy of 68% has been reported which gradually increases to 91%.

1.1. Related Works

In [23], a scheme has been proposed for mining association rules using principal component analysis (PCA). This is exclusively for cricket matches. They have proposed a framework for establishing correlations between pieces of cricket statistics with frequent patterns. This framework is meant to help in making and improving coaching strategies. In [24], the same association rule mining has been implemented for strategic planning for teams during ICC-2015. Several decisive parameters like match-venue, toss output, rank order of a batsman, strike-rate, and score-economy has been analyzed. In [25], performance data mining has been presented for the cricket team of New-Zealand. It takes into consideration all the historical data pertaining to the New-Zealand versus other teams starting from the year 1975. In [26], a combined approach of few of the modern classification techniques has been analyzed for the prediction of ODI cricket outcomes. Naive Bayesian, Support Vector Machines (SVM), and Random Forest (RF) have been used for the purpose.

2. Motivation and Objectives

So far, numerous techniques have been proposed for several aspects in cricket. However, there remain the limitations that are need to be addressed with efficient heuristics. It is learned from the literature that, for the two different format of the game of cricket, a single prediction strategy may not yield fruitful prediction outcome. This is because of the fact, the match being played by the teams in these two formats are with strategically distinct approaches. Thus, there has been a need for two distinct schemes for two of this formats (ODI and T-20) of the same game. In this work, such an attempt has been made to propose the winning probability for a team for the two formats of a cricket match (ODI and T-20). The objective of this work is to devise suitable prediction strategies for the ODI and T-20 versions of the cricket game separately. Optimality need to be given utmost priority while designing the prediction strategies. The organization of this paper is as follows. In the subsequent sections the background, proposed heuristics, and experimental analysis have been illustrated in a sequence. Final conclusion has been made in the last section.

3. Background

The background characteristics are depicted as under.

- The game of cricket is played between two teams, with each team having eleven numbers of live players.
- A toss is performed at the beginning for the act of choosing one of the option from *fielding and bowling and batting* as the first choice by a team,
- For a team, the game result can be a win, or a loss, or a draw. (Let's not consider the scenario of matches getting abandoned),
- For the bowling team, an over refers to act of delivering the ball six times as per the rule of bowling actions. Invalid action of delivering the ball may be considered as extra runs in terms of wide-ball or no-ball which may be awarded to the opponent team (batting team).
- The score made by a batting team has to be chased by the opponent team. Upon completing the chase successfully, the team is declared winner, else a loosing team.
- A one-day-cricket is played for a single day. this comprises of a total of a hundred of overs with individual team acting for bowling and batting for fifty overs each.
- A T-20 match, as the name suggests, comprises of a total of forty overs with individual team acting for bowling and batting for twenty overs each.

4. Proposed Scheme

The proposed scheme comprises of two different approached for the two different formats of the game. These have discussed in a sequel.

4.1. Approach-1(One-day-cricket)

This is a statistical approach which considers the data pertaining to the team's performance and individual players performance in the recent past. The direct way of doing the analysis and prediction is to verdict for the team with overall good performance (GP). The GP can be defined as the sum of individual computations as given below:-

$$GP = \left[\frac{tot\_won}{tot\_played} * 100\right] + \left[\frac{tot\_won\_vs}{tot\_played\_vs} * 100\right] + \sum_{i=1}^{11} IP_i * 100 \tag{1}$$

where,

- tot_won is the total number of matches won by the team till date,
- tot_played is the total number of matches played by the team till date,
- tot_won_vs is the total number of matches won by the team versus the current opponent,
- tot_played_vs is the total number of matches won by the team versus the current opponent,
- IP_i is the individual performance of the players of the team. This can be computed based on the formula as given in the equation below:-

  $$IP_i = BP_i + WP_i \tag{2}$$

  where, BP_i and WP_i are the batting performance and bowling performance of a player respectively. These parameters are given by the formula as described below:-

  $$BP_i = w_1 * \left[\frac{\#fifty}{\#matches}\right] + w_2 * \left[\frac{\#hundred}{\#matches}\right] + w_3 * \left[\frac{sr}{100}\right] \tag{3}$$

  where, w_1, w_2, and w_3 are the weight values which can be computed using the function as given below,

  $$w_1 = \frac{\text{NFL}}{\text{NFS}} \tag{4}$$

  $$w_2 = \frac{\text{NHL}}{\text{NHS}} \tag{5}$$

  $$w_3 = \frac{\text{SR}}{\text{TSR}} \tag{6}$$

  $$WP_i = \left[\frac{\text{NWT}}{tot\_played}\right] + \left[\frac{\text{NWF}}{tot\_played}\right] - penalty \tag{7}$$

  where, $penalty$ = The penalty due to extra runs given through *no ball*, and *wide ball*. This can be computed as,

  $$penalty = \frac{extra}{delivered \times 6.0} \tag{8}$$

where,

- NFL is the number of fifty runs scored in last ten matches,
- NFS is the number of fifty runs scored so far,
- NHL is the number of hundred runs scored in last ten matches,
- NHS is the number of hundred runs scored so far,
- SR is the strike rate for the last ten matches,
- TSR is the total strike rate in career,
- NWT is the number of 3-wickets taken in last ten matches,
  item NWF is the number of 5-wickets taken in last ten matches,
- $extra$ is the total number of extra balls delivered to the current opponent,
- $delivered$ is the total number of extra balls delivered so far in career,
- The constant value 6.0 has been introduced in the denominator because there are six valid balls delivered in an over.

4.2. Approach-2(T-20)

This approach takes into consideration the aggression factor. This is because, a T-20 format cricket involves a match of a total of 40 overs, where each of the two teams are allowed to bat/ bowl for 20 overs. That makes a sense of aggression towards the teams. As the number of overs are less, hence the players individual performance has to consider an aggression factor (α for batsman and β for bowler). Thus the equations for computing the overall performance of a team also needs to be altered. The corresponding set of equations for the same computation are given as under.

$$GP = \left[\frac{tot\_won}{tot\_played} * 100\right] + \left[\frac{tot\_won\_vs}{tot\_played\_vs} * 100\right] + \sum_{i=1}^{11} IP_i * 100 \tag{9}$$

$$IP_i = \alpha * BP_i + \beta * WP_i \tag{10}$$

where, α and β are the aggression factors. These can be assumed to be probabilistic values as they are somehow dependent on the outcome of a toss. The toss is essentially important for predicting the winner so far as a T-20 match is concerned. Another influential factor is the pitch condition on which the match would be played. Hence, these two parameters are taken as per the formula given below:

$$\alpha = \begin{cases} 0.5, & \text{if toss is won} \\ 0.1, & \text{otherwise} \end{cases} \tag{11}$$

$$\beta = \begin{cases} 0.25, & \text{if pitch is wet} \\ 0.125, & \text{otherwise} \end{cases} \tag{12}$$

4.3. Prediction Probability

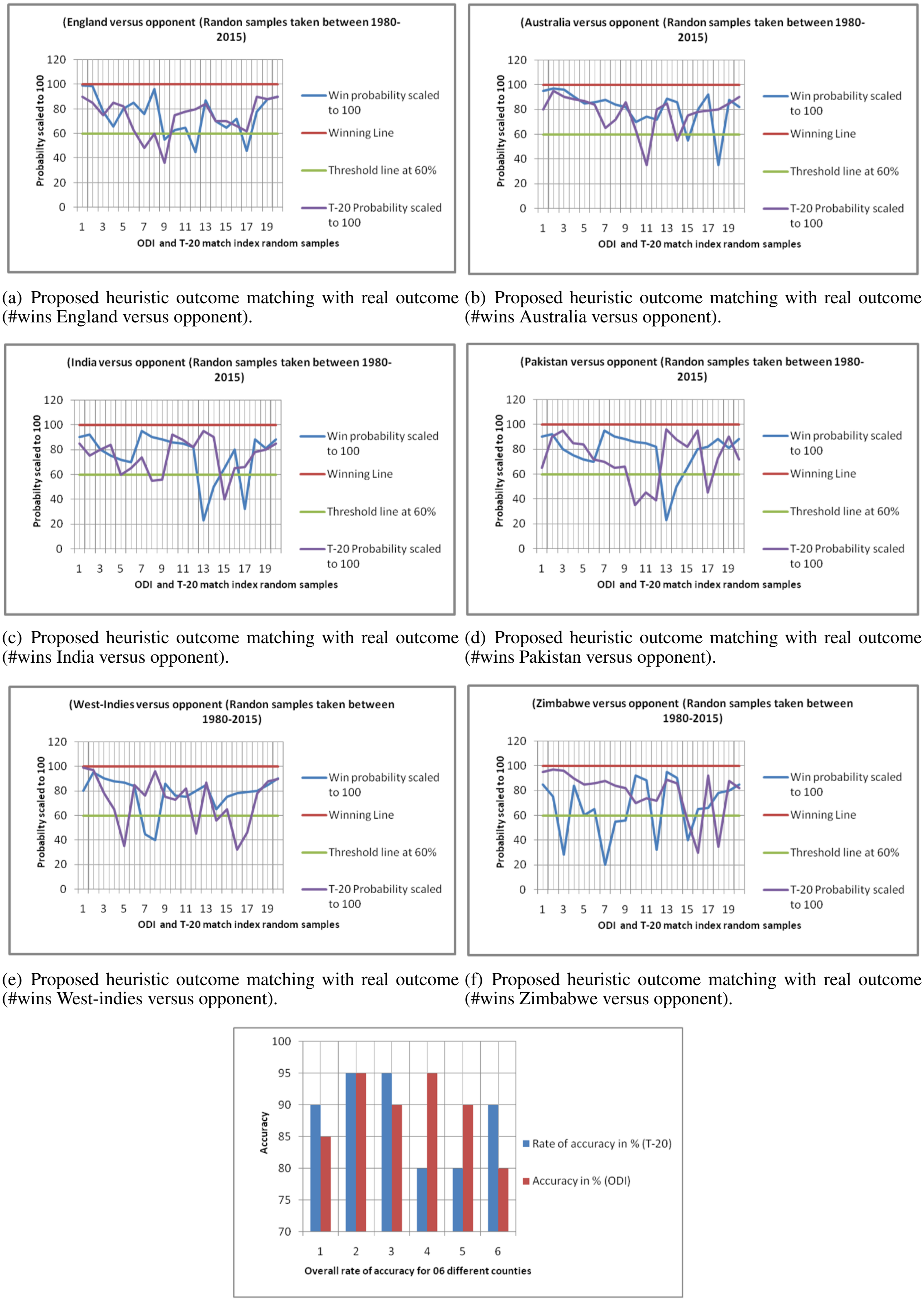
For each of the two approaches discussed in the previous sections, the prediction probability can now be found as per the formula as given in the equation below:

$$P(win) = (\alpha + \beta) * \left[1 - \frac{1}{GP}\right] \tag{13}$$

The parameters α and β have been considered for the final probability calculation. This is because the overall outcome of the match somehow considered to dependent on those irrespective of the version type of the cricket format.

5. Case Study

The proposed heuristics are considered for case studies on the data of performance of six distinct countries. The data considered here are for both the ODI and T-20 format of matches. Random samples of winning matches for the teams are selected along with mentioned attributes and the proposed schemes are applied individually. The prediction outcomes are obtained in terms of probabilistic values. These values are scaled to a value of 100 and are plotted against the real match outcomes of those corresponding matches. For the obtained probabilistic results, a threshold is set as 60% so that it can be considered valid if and only if it is above that threshold. The threshold percentage is set to be 60% instead of 50% only to assume a fare estimation of the efficiency of the schemes. The dataset available in *www.kaggle.com* has been considered for the purpose. The said website draws the dataset from the espn private limited which is the leading sports media company. Kaggle is a genuine and popular machine learning quest platform that facilitate vivid variety of challenges for researchers for grooming their competencies. The overview of the dataset particularly used for this case-study has been depicted in Table 1.



(a) Proposed heuristic outcome matching with real outcome (#wins England versus opponent).

(b) Proposed heuristic outcome matching with real outcome (#wins Australia versus opponent).

(c) Proposed heuristic outcome matching with real outcome (#wins India versus opponent).

(d) Proposed heuristic outcome matching with real outcome (#wins Pakistan versus opponent).

(e) Proposed heuristic outcome matching with real outcome (#wins West-indies versus opponent).

(f) Proposed heuristic outcome matching with real outcome (#wins Zimbabwe versus opponent).

(g) Overall rate of accuracy for 06 counties.

In Figure1(a), it can be observed that, both of the ODI and T-20 predictions are most often falling above the threshold percentage. This shows that the scheme is satisfactorily working with rate of accuracies of 85% and 90% for both of the match formats. Similarly, the prediction plots for both of the formats are plotted for five other countries namely, Australia, India, Pakistan, West-indies, and Zimbabwe in Figures 1(b) - 1(f) respectively. All these plots show similar satisfactory rates of accuracies. This makes the proposed schemes appear to be robust and efficient. The overall rate of accuracy for each of the six teams has been shown in Figure 1(g), where both the values of accuracies in ODI and T-2- formats have been presented. The rates are in the range of 80-95 % which is satisfactory. A comparative analysis of the proposed schemes have been made with state-of-the-art machine learning (ML) schemes. These schemes have been simulated on the same set of samples and the results so obtained are compared with the proposed scheme. This comparison has been shown in Table 2. The proposed scheme outperforms the rest with a sufficiently good marginal difference.

Table 1. Overview of the dataset used.

Title	**Description**
Name:	Cricket ODI and T-20 dataset
Source	www.kaggle.com
Total teams considered	06(six)
#of base attributes	08(eight)
Number of derived attributes	06(six)
Duration	**1980-2015**

Table 2. Performance comparison with state-of-the-art ML tools.

SL#	**Tool/ Algorithm**	**Data**	**Accuracy (%) ODI**	**Accuracy (%) T-20**
1	Random forest	Historical	78	76
2	Regression	Historical	65	54
3	Naive Baye's	Historical	76	79.5
4	MLP	Historical	58	62
4	SVM	Historical	82	82
5	Proposed heuristics	Historical	89.1	88

6. Conclusion

Heuristics for efficient prediction of the winner for a cricket match of two different formats have been proposed. The proposed approaches consider every important aspects which are directly and indirectly effecting the outcome of a match. Here, statistical data are used to derive at a concluding single parametric value which is finally used in a suitably defined probabilistic function for predicting the winning probability of a team. These approaches are unique of their kind as they do not incorporate any type of predefined classifiers. Test cases have been considered from benchmark dataset for evaluation purpose. Further, these methods upon comparison with other schemes those using benchmark classifiers give comparatively better performance in terms of the overall rate of accuracies (89.1% for ODI and 88.33% for T-20). The future work may focus on devising a dynamic approach for live match prediction by taking the outcomes of these methods as a prior.

REFERENCES

1. Al-Zahrani and M A Ali, *Recurrence Relations for Moments of Order Statistics from the Lindley Distribution with General Multiply Type-II Censored Sample Bander*, Statistics, Optimization & Information Computing, Vol. 2, no. 2, pp. 147 - 160, 2014.
2. C. Parpoula, C. Koukouvinos, D.E. Simos and S. Stylianou, *Supersaturated plans for variable selection in large databases*, Statistics, Optimization & Information Computing, Vol. 2, pp. 161 - 175, 2014.
3. Noryanti Muhammad, Tahani Coolen-Maturi, Frank P.A. Coolen, *Nonparametric predictive inference with parametric copulas for combining bivariate diagnostic tests*, Statistics, Optimization & Information Computing, Vol. 6, pp 398C408, 2018.
4. M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, *An analytical method for diseases prediction using machine learning techniques*, Computers & Chemical Engineering, vol. 106, pp. 212 C 223, 2017.
5. S. A and A. T. T, *Prediction of heart disease complication for diabetic patient using data mining techniques*, International Journal of Pure and Applied Mathematics, pp. 13 869C13 879, 119 2018.
6. F. Kunihiko, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological Cybernetics, vol. 36, no. 04, pp. 193C202, 1980.
7. Shubham Agarwal, Lavish Yadav and Shikha Mehta, *Cricket Team Prediction with Hadoop: Statistical Modeling Approach*, Procedia Computer Science, vol. 122, pp. 525 - 532, 2017.
8. Subramanian Rama Iyer and Ramesh Sharda, *Prediction of athletes performance using neural networks: An application in cricket team selection*, Expert Systems with Applications, vol. 36, no. 3, Part 1, pp. 5510 - 5522, 2009.
9. Rajitha M. Silva, Ananda B.W. Manage and Tim B. Swartz, *A study of the powerplay in one-day cricket*, European Journal of Operational Research, vol. 244, no. 3, pp. 931 - 938, 2015.
10. Neeraj Pathak and Hardik Wadhwa, *Applications of Modern Classification Techniques to Predict the Outcome of ODI Cricket*, Procedia Computer Science, vol. 244, pp. 55 - 60, 2016.
11. Muhammad Asif and Ian G. McHale, *In-play forecasting of win probability in One-Day International cricket: A dynamic logistic regression model*, International Journal of Forecasting, vol. 32, no. 1, pp. 34 - 43, 2016.
12. Hugh Norton, Steve Gray and Robert Faff, *Yes, one-day international cricket in-play trading strategies can be profitable!*, Journal of Banking & Finance, vol. 61, pp. S164 - S176, 2015.
13. Sohail Akhtar and Philip Scarf, *Forecasting test cricket match outcomes in play*, International Journal of Forecasting, vol. 28, no. 3, pp. 632 - 643, 2012.
14. H. Ahmad, A. Daud, L. Wang, H. Hong, H. Dawood and Y. Yang, *Prediction of Rising Stars in the Game of Cricket*, IEEE Access, vol. 5, pp. 4104 - 4124, 2017.
15. T. Singh, V. Singla and P. Bhatia, *Score and winning prediction in cricket through data mining*, 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), pp. 60 - 66, 2015.
16. *http://www.cricwaves.com/*
17. *http://www.crickbuzz.com/*
18. *https://fantasycricket.dream11.com/in/*
19. *https://www.espncricinfo.com/*
20. Carson K. Leung, Kyle W. Joseph, *Sports data mining: predicting results for the college football games*, 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES 2014, pp. 710 - 719, 2014.
21. A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan and I. Matthews, *Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data*, 2014 IEEE International Conference on Data Mining, pp. 725 - 730, 2014.
22. H. Janetzko, D. Sacha, M. Stein and T. Schreck, D. A. Keim and O. Deussen, *Feature-driven visual analytics of soccer data*, 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 13 - 22, 2014.
23. P. UmaMaheswari and M. Rajaram, *A Novel Approach for Mining Association Rules on Sports Data using Principal Component Analysis: For Cricket match perspective*, 2009 IEEE International Advance Computing Conference, pp. 1074 - 1080, 2009.
24. S. Bhattacherjee, J. Sahoo and A. Goswami, *Association Rule Mining Approach in Strategy Planning for Team India in ICC World Cup 2015*, 2015 Second International Conference on Advances in Computing and Communication Engineering, pp. 616 - 621, 2015.
25. R. K. Khan, I. Manarvi and Mohay-ud-din, *Evaluating performance of Blackcaps of New Zealand vs. global cricket teams*, 2009 International Conference on Computers Industrial Engineering, pp. 1500 - 1504, 2009.
26. Neeraj Pathak, Hardik Wadhwa, *Applications of Modern Classification Techniques to Predict the Outcome of ODI Cricket*, Procedia Computer Science, no. 87, pp. 55 - 60, 2016.
27. R. Bryll, R. Gutierrez-Osuna, and F. Quek, *Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets*, Pattern Recognition, Vol. 36, no. 6, pp. 1291 - 1302, 2003.
28. *http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/similarity.html*
29. Szekely, G.J. and Rizzo, M.L., *Data mining and knowledge discovery, Springer, The Netherlands*, The Annals of Statistics, Vol. 42, no. 6, pp. 121 - 167, 2014.