This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

Prediction of Rising Stars in the Game of Cricket

Haseeb Ahmad, Ali Daud, Licheng Wang, Haibo Hong, Hussain Dawood, and Yixian Yang

***Abstract*—Online social databases are rich sources to retrieve appropriate information that is subsequently analyzed for forthcoming trends prediction. In this work, we identify rising stars in cricket domain by employing machine learning techniques. More precisely, we predict rising stars from batting as well as from bowling realms. For this intent, the concepts of Co-players, Team and Opposite teams are incorporated and distinct features along with their mathematical formulations are presented. For classification purpose, generative and discriminative machine learning algorithms are employed, and two models from each category are evaluated. As a proof of applicability, the proposed approach is validated experimentally, while analyzing the impact of individual features. Besides, model and category wise assessment is also performed. Employing cross-validation, we demonstrate high accuracy for rising star prediction that is both robust and statistically significant. Finally, ranking lists of top ten rising cricketers based on weighted average, performance evolution and rising star scores are compared with the international cricket council rankings.**

***Index Terms*—Cricket, machine learning, online social databases, prediction, rising stars.**

I. INTRODUCTION

RISING Stars (RSs) are ones, who currently own relatively weak profiles , but can be predicted as the future experts of the respective domains [1]. Rising Star Prediction (RSP) is made based on the current contributions of RSs coupled by considering their ascending performance and joint collaborations with the domain experts. Finding such RSs within the organizational domains is the great need of current era, so that the organizations can put efforts to maximize the expertise of RSs in order to get the optimal performance in future. Hence, RSP is an emerging research dimension that inspires us to forecast RSs from sports domain. Considering the sports domain, among all sports, cricket is the second most popular game that was originated from England, and now has its roots round the globe. Therefore, we take cricket game as our case study for RSP. The players in cricket game can be categorized into different classes based on their performance evolution. Although there could be many evolution classes, but usually a player belongs to any of the four evolution classes that are presented in Figure 1. Here, it is necessary to distinguish between a star and a rising star. More precisely, a star in cricket is an experienced player with extraordinary performance throughout his running career. On the other hand, a rising cricketer or a rising star is an emerging player, who currently has a low profile, but could be a star cricketer in future based on consistent improvements in performance. Hence, these RSs arise as foremost contributors to their squads. Yet, there is no existing criteria to extract emerging batsmen and bowlers in cricket game. Thus, searching such RSs is a new dimension that inspires us to extract exceptional rising cricketers in sports domain.

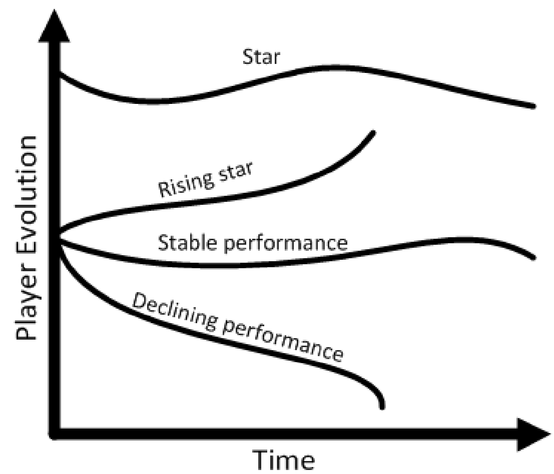


Fig. 1. Four main evolution stages of sports players.

Searching, ranking and prediction of domain experts or contents has gained the attention of the researchers with the advent of online social networks and online databases. Several works are put forward by scholastic community to address these issues. Finding RSs in academic networks [1], [2], future prediction of citation count [3] and temporal expert finding [4] are such proposal among many other nuanced results. But these proposals did not consider sports domain. While considering the sports domain, some issues related to sports forecasting are discussed in [5], but proper prediction mechanism is not presented by the authors. Moreover, none among above proposals considered cricket domain. Taking cricket game into an account, David, and Swartz extracted best batsmen and bowlers [6]. There are further proposals for ranking of batsman based on performance metric [7] and for ranking of cricket teams by employing h-index and PageRank [8]. However, the aforementioned proposals did not put forward any mechanism for RSP in the cricket domain.

In this work, for RSP, we put forward the concept of Co-players (including stars, rising stars and stable performers), who have played with an emerging batsman or a bowler

Haseeb Ahmad, Licheng Wang and Yixian Yang are from State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China. (e-mail: haseeb_ad@hotmail.com; wanglc@bupt.edu.cn; yxyang@bupt.edu.cn).

Ali Daud is from Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, and Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan.(e-mail:adfmohamad@kau.edu.sa)

Haibo Hong is from School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China.(e-mail:honghaibo1985@163.com)

Hussain Dawood is from Faculty of Computing and Information Technology,University of Jeddah, Jeddah, Saudi Arabia.(e-mail:hdaoud@uj.edu.sa)

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

in joint ventures under same playing conditions (i.e. faced same opposition, home or away grounds, fast or slow pitches, hot or cold weather etc.). Thus, we formulate a criteria for RSP in such a way that the early years performance of an emerging player and the performances of its Co-players are incorporated. In this manner, an emerging player finds the opportunity to learn from the playing strategies of Co-players under the same playing conditions in order to improve its performance. Obviously, to be included in the same team, an emerging player has to be more productive while learning from Co-players and competing with them at the same time. Otherwise, the underlying player will be considered as a declining performer and would be excluded from team. Thus, the concept of Co-players can be considered as an essential factor to be incorporated for RSP in cricket domain. For this purpose, we define three categories of features (Co-batsmen, Team and Opposite teams) for batsmen as well as similar categories (Co-bowlers, Team and Opposite teams) for bowlers. These modules properly integrate the concept of Co-players. Furthermore, several features are inquired under these categories and their mathematical notions are derived. Moreover, machine learning based classification models (generative and discriminative) are employed for RSP in cricket game. Outcome of each classifier addresses the hypothesis question: "H = Underlying emerging player is probable to come forth as a future cricket star or not?" The major contributions of our work are briefed as follows:

1) We present an efficient methodology for RSP within the cricket domain while incorporating the concept of Co-players. A set of 9 features is formulated for RSP of batsmen as well as a set of 11 features is scrutinized for the bowlers. These features are never considered before.
2) By testing different classification algorithms on our datasets, four appropriate machine learning classification algorithms are selected for binary classification of RSs.
3) The performance of employed machine learning algorithms is critically examined during the evaluation phase. It is found that generative classifiers outperforms the others.
4) RSP is made with high accuracy, and rankings of leading RSs from both domains based on three defined metrics are presented and compared with the ICC rankings of players from 2013-2016.
5) This innovative idea can be used for RSP in other sports domains such as baseball, football and basketball etc.

A. Related Work

Online social networks including Twitter, Facebook and databases like dblp[1], ESPNCricinfo[2], Arnetminer[3] have revolutionized the globe. Along with abundant information and entertaining facilities, these networks have also become worthwhile sources for marketing and got the temptations as a research domain to expedite more innovative services. In 2009, Li et al. proposed PubRank algorithm to mine RSs, but it has limitation as it incorporates only static ranking of venues and author mutual influence [1]. A subsequent work put forward by Tsatsaronis et al. suggested a model to address the dynamics of authors' bibliography [9]. Unsupervised learning procedures were employed to categorize the authors into different groups. Another work towards finding future talent was presented by Daud et al., in which the authors improved PubRank while incorporating mutual influence of authors and venue scores [10]. In a recent work, Daud et al. utilized supervised machine learning classification models for RSP in co-author networks [2]. There are further proposals, in which discriminative (Support Vector Machines (SVM), Classification And Regression Tree (CART)) [11], [12] were employed for classification and analysis purposes. Similarly, generative classifiers (Bayesian Network (BN), Naïve Bayesian (NB)) were also used for prediction and classification purposes [13], [14]. However, none of the previous proposals provided a mechanism for RSP in cricket.

Nevertheless, very little efforts are made towards providing efficient mechanisms for ranking teams, or to find experts from the teams. Recently, some social networking analysts among scholastic communities have attracted towards this dimension due to its over grown popularity. Working towards this direction, Bracewell & Ruggiero utilized parametric control chart for performance monitoring of a batsman in different matches [15]. Amin & Sharma employed ordered weighted averaging technique coupled with regression structure in order to measure the performance, and to rank T20 batsmen [7]. A different approach was put forward by Mukherjee [16], in which state-of-the-art PageRank [17] technique was suggested for team and captain rankings. Although, this was the first effort of applying PageRank in cricket domain for ranking purpose, but the author ignored the role of runs and wickets by which the winning teams gain victory. Daud et al. noticed this shortcoming and proposed the hybridization of graph and non-graph weightage model for ranking the cricket teams [8]. In another work, Mukherjee put forward a gradient network based approach to incorporate the concept of Co-players [18]. This strategy nominates the batsmen as Co-players, who face the same bowlers, or the bowlers who bowled to the same batsmen. Subsequently, these players are ranked by PageRank algorithm. Although the concept of Co-players was introduced, but the playing conditions such as home/away grounds, fast/slow pitches etc. were ignored. Moreover, author only used batting and bowling averages to rank the players and many other important metrics such as strike rate, win/loss ratio etc. were ignored. Another work towards forecasting about upcoming test match session result based on logistic regression model was presented by Akhtar and Scarf [19]. Besides such analytics in this field, none of aforementioned proposals considered the mining of RSs from the cricket domain. Therefore, we are motivated to provide an efficient mechanism to address this gap for the betterment of cricket game.

The remaining contents of paper are structured as follows. Section II describes the basic concepts and terminologies including a brief introduction of cricket game, its rules and regulations and ranking metrics. Besides, Section III highlights

[1]http://dblp.uni-trier.de/
[2]http://www.espncricinfo.com/
[3]https://cn.aminer.org/

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

the problem definition. Section IV reviews the machine learning models adopted for evaluation. In Section V, we present the concept of Co-players and introduce its features. Additionally, this section provides the mathematical notions of defined features. Experiments and evaluations are discussed in Section VI. Section VII provides the ranking of RSs based on three different scores, while Section VIII details the concluding thoughts.

II. Basic Concepts and Terminologies

A. Cricket Game

Cricket is a sport that is played in a ground between two teams with a bat and ball, where each team is comprised of 11 players. Both of the teams act as opponent and try to win the match against the other. Each team is required to bat during its turn, while the opposite team has to bowl and field in the ground during this phase.

1) International Cricket Council: International Cricket Council (ICC)[4] is the governing authority, responsible for making and amending the rules for cricket game. Moreover, the cricket teams and players rankings are also issued by ICC on regular basis.

2) Team Structure: The four main roles of each team are categorized as batsman, bowler, wicket keeper and all-rounder. Further, there is a designation of captain in each team. Captain is nominated by the authorities of each team and it could be anyone among the playing eleven. Before starting each match, a coin is tossed. The toss winning captain decides whether they have to bowl or bat first. As the match starts, the bowler bowls to the opposite team's batsman to get him out, while the batsman tries to hit the ball in order to gain maximum runs. During this process wicket keeper fields behind the batting wicket and the remaining players from the bowling side also field with in the ground boundaries. All-rounder is the player who acts as a batsman and as well as a bowler.

3) Basic Rules of Cricket: A circle shaped field that hosts the cricket match is called a ground. The size of ground varies from 65 meters to 95 meters. In the center of ground, a rectangular shaped 22 yards long zone is called *pitch*. An over is comprised of six balls that are bowled consecutively by a single bowler. Bowler bowls the ball to the batsman of opposite team on pitch and batsman tries to hit the ball for getting scores. The batsman can score one to six runs on one ball. When the batting turn of a team ends, the opposite team starts batting for chasing the total runs made by the first team. Each turn is known as an innings. If the latter batting team successfully scores the total runs within limited overs (one day match) with the wickets in hands, it wins, otherwise the former batting team wins with the number of un-chaseable score. This game seems similar to that of baseball, but the rules and regulations are different.

4) Cricket Game Formats: There are a variety of cricket formats including 1) Test Match; 2) One Day International (ODI); and 3) Twenty20 (T20). Test match is played for consecutive five days, in which each team has to play two innings. Each team during every innings usually continues to make score until ten batting players are got out by the bowling team. Normally, a maximum of 90 overs are played each day. However, there is no restriction for any team to play limited overs during each innings. ODI is a faster format as compared to test match. Each ODI is completed in one day, while each team is restricted to bat for 50 overs unless ten players get out earlier. T20 is the fastest format of known time. Each team is restricted to bat for 20 overs only in a single innings.

5) Ranking Metrics: Based on the performance of each team and player (batsman/bowler), ICC issues the rankings on regular basis. Each team gets or loses points while winning or losing cricket matches, respectively. Considering a particular time span, these points are utilized for ranking the teams by ICC. Some of basic metrics for ranking the batsmen are listed as: 1) runs (total number of scores made by a batsman); 2) batting average (number of runs made by the batsman divided by number of times it got out during career span); and 3) batting strike rate (number of scores made by the batsman per 100 balls faced). Similarly, the metrics for ranking bowler are: 1) bowling average (numbers of runs conceded by the bowler divided by the number of wickets taken); 2) bowling strike rate (number of balls bowled by the bowler divided by number of wickets taken); and 3) bowling economy (number of runs conceded by the bowler per over). These metrics are employed in the later section while defining features for RSP.

III. PROBLEM DEFINITION

The usual evaluation of a cricketer is assessed only by considering his personal performance. The features such as collaboration with star cricketers, overall team performance, opposite teams performance, win/loss ratio etc. are never taken into account for judging the performance of an emerging cricketer. Obviously, these features along with many others also help for grooming the abilities of a RS. Consequently, performance checks while including these metrics can provide improved results as compared to ordinary assessments. Based on these more effective results, RSP can be made more precise. Next, we present the formal description of RSP in cricket game.

A. Rising Star Prediction

Given a set of tuples with n training examples $(\mathbf{X}_1, y_i), (\mathbf{X}_2, y_i), \ldots, (\mathbf{X}_n, y_i)$, where, $\mathbf{X}_i$ denotes the feature vector of cricketer c_i, while $\mathbf{X}_i \in R^m$, R is the real features space, m is the total count of features, n is the total count of underlying cricketers and $y_i \in \{-1, +1\}$. Moreover, for RSP a prediction function P_{RS} is defined as follow [2]:

$$F = P_{RS}(c_i/\mathbf{X}), \tag{1}$$

where,

$$P_{RS}(c_i/\mathbf{X}) = \begin{cases} < 0 & \text{if } y = \text{-1}, \qquad notRS, \\ \geqslant 0 & \text{if } y = \text{+1}, \qquad RS. \end{cases} \tag{2}$$

[4]http://www.icc-cricket.com/home

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

B. Objective

The objective is RSP, i.e. it is required to learn whether a cricketer is a RS or not after a time span Δt. Formally, we need to inquire a predictive function $\hat{P}_{RS}$ as follows:

$$\hat{F} = \hat{P}_{RS}(c_i/\mathbf{X}, \ \ \Delta t). \tag{3}$$

We have extracted many important features that are useful for RSP. For each cricketer, these features predict the anticipated label accurately. We provide the mathematical structures of applied classifiers in the next section.

IV. MODELS

The swift progression of scientific research has made it possible to predict the foreseeable future of almost every potential domain. These predictions are usually made by employing scientific tools such as machine learning classification models and stochastic models etc. Classification models are further divided into two classes. 1) Generative (that randomly generate the observable data values while given some latent variables), and 2) discriminative (that model the dependence of unknown variable over known variable). Next, we provide a brief introduction of each model for learning the predictive function $\hat{P}_{RS}$.

A. Generative Models

We review important characteristics for central comparative analysis of two widespread generative classifiers: BN and NB in this section.

1) Bayesian Network (BN): A BN is a directed acyclic graph representing a joint probability distribution over a set of random variables in terms of their conditional dependencies [13]. BN is a robust classifier as small changes in model do not result in dramatic affects. Moreover, BN can be used for prediction, classification and configuration problems. The joint probability density function of a BN comprised of n nodes $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$is given as follows:

$$\begin{aligned} P(\mathbf{X}_1 = x_1, \ldots, \mathbf{X}_n = x_n) &= P(x_1, \ldots, x_n) \\ &= \prod_i P(x_k | x_1, \ldots, x_{i-1}) \end{aligned} \tag{4}$$

In addition, all the nodes within BN are conditionally dependent on their parent nodes. Thus,

$$P(\mathbf{X}_1 = x_1, \ldots, \mathbf{X}_n = x_n) = \prod_k P(\mathbf{X}_k | Parents(\mathbf{X}_k)) \tag{5}$$

For binary classification using BN, first we need to calculate the conditional probabilities for each class as follows:

$$P(y_i|\mathbf{X}) = \frac{\prod_k P(\mathbf{X}_k | Parents(\mathbf{X}_k))}{P(\mathbf{X})}, \ \ where \ y_i \in \{-1, +1\}. \tag{6}$$

Finally, BN assigns $\mathbf{X}_k$ to the class with $max_y P(y_i|\mathbf{X})$.

2) Naïve Bayesian (NB): NB is the first successor classifier of BN, but with additional difference of independence between the features [14]. NB classifiers offer high scalability as these can handle both categorical independent as well as continuous variables. Moreover, NB also entertain small number of instances for the estimation of necessary parameters that are required for classification.

Given a set of feature vectors $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ and the class label y, NB classifier will assign $\mathbf{X}$ to the class having maximum posteriori probability that is given as follows:

$$P(y_i|\mathbf{X}) = \frac{P(\mathbf{X}|y_i)P(y_i)}{P(\mathbf{X})}, \ \ where \ y_i \in \{-1, +1\}. \tag{7}$$

Here, it should be noted that P($\mathbf{X}$) remains the same for all classes, therefore, we need to maximize $P(\mathbf{X}|y_i)P(y_i)$. Further, $P(y_i)$ also remains the same as there are equal number of instances for both classes. Thus, we are only required to maximize $P(\mathbf{X}|y_i)$, which is an expensive task in terms of computations. By embedding assumption of conditional independence (Naïve) approach, computational cost is reduced. Conditional independence presumes that attributes values are independent from each other, provided the class labels. Hence, the expression can be reduced to follows:

$$P(\mathbf{X}|y_i) \approx \prod_{k=1}^{n} P(\mathbf{X}_k|y_i). \tag{8}$$

Now, $P(\mathbf{X}_k|y_i) = P(\mathbf{X}_1|y_i), P(\mathbf{X}_2|y_i), \ldots, P(\mathbf{X}_n|y_i)$ can be easily calculated for each class and subsequently, $\mathbf{X}$ is assigned to the class having maximum posteriori probability. For instance, $\mathbf{X}$ will be assigned to RS class $(y_{(+1)})$ iff following holds.

$$P(y_{(+1)}|\mathbf{X}_i) > P(y_{(-1)}|\mathbf{X}_i). \tag{9}$$

B. Discriminative Models

We concentrate on reviewing models of SVM and CART with the resolution of debating fundamental issues in an important comparative analysis of discriminative class.

1) Support Vector Machines (SVM): Among state-of-the-art binary classifiers based on supervised machine learning, Support Vector Machines (SVM) have gained broader popularity due to efficient investigation of data while identifying the patterns [11]. More precisely, for efficient separation of two different classes, SVM model constructs the optimal hyperplane with largest functional margin. Moreover, it can handle linear and non-linear data.

Given a set of tuples with n training examples $((\mathbf{X}_1, y_i), (\mathbf{X}_2, y_i), \ldots, (\mathbf{X}_n, y_i))$, where $y_i \in \{-1, +1\}$, each y_i points out to which class (not RS, or RS) the corresponding feature vector $\mathbf{X}_i$ belongs to. The intuition here is to construct a hyperplane with maximum functional margin that can divide a group of feature vectors $\mathbf{X}_i$ with $y_i = +1$ (RS) from that of with $y_i = -1$ (not RS) while minimizing the classification error. The hyperplane with set of points $\mathbf{X}$ satisfies the following equation:

$$\mathbf{w}^T \cdot \mathbf{X}_k + b = 0 \tag{10}$$

where, $\mathbf{w}$ is the vector, normal to the hyperplane and b is the offset. Both $\mathbf{w}$ and b are calculated during SVM training. Furthermore, the offset of the hyperplane from the origin along $\mathbf{w}$ is regulated by the factor $b/(||\mathbf{w}||)$. Thus, we have to solve the following primal problem to get the solution of binary

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

classifier that could separate *RSs* and *not RSs*.

$$min(||\mathbf{w}||), \qquad subject\ to \quad y_i(\mathbf{w}^T \cdot \mathbf{X}_k + b) \geq 1, \quad for\ k = 1, \cdots, n. \tag{11}$$

2) Classification And Regression Tree (CART): CART is fundamentally a non-parametric model used for making prediction on underlying data. Basically, CART is comprised of three main steps [12].

1) Maximum tree construction
2) Right selection of tree size
3) Classification of unseen data based on former trained tree

Given input feature vector $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, our goal is to forecast a feedback of class y. Thus, a binary tree is set up while performing a test on each non-leaf nodes to construct a right or left sub-branch of tree. The development of binary tree is continued until leaf nodes are constructed. To check the impurity of each node, Gini index is utlized as follows:

$$Gini\ index(t_p) = I(t_P) = 1 - (\frac{n_{(+1)}}{n})^2 - (\frac{n_{(-1)}}{n})^2, \tag{12}$$

where, $I(t_p)$ is the impurity measuring function of under consideration node t_P, (+1) and (-1) are two classes depicting *rising stars* and *not rising stars*, respectively. $n_{(+1)}$ and $n_{(-1)}$ are denoting the numbers of subjects present at t_P, which belong either to RS or Not RS class, respectively. n refers to total number of subjects present at t_P. More precisely, at each node, CART resolves the following optimization problem.

$$arg\ max_{\mathbf{X}_i \leq \mathbf{X}_i^R, i=1,\ldots,n}[I(t_P) - P_L\ I(t_l) - P_R(t_r)] \tag{13}$$

where, t_P, t_r and t_l denote parent, right child and left child nodes, respectively. While P_R and P_L are the probabilities of right and left child nodes, respectively.
In details, CART produces a sequence of sub-trees for developing a large tree and subsequently, prune it backward until only root node is left. Afterward, for the estimation of misclassification cost of each subtree, it employs cross-validation and finally selects the one having the lowest cost.

C. Feature Evaluation Metrics

Three state-of-the-art feature evaluators including *information gain*, *gain ratio* and *chi-squared statistic* are employed for mining effective features for ranking and prediction tasks [20]. The mathematical notions of these metrics are presented as follows:

1) Information Gain: In our scenario, information gain of a feature $\mathbf{X}_k$ while classifying an emerging player as $y_i \in \{-1, +1\}$ (RS or Not RS) is presented as follows:

$$Info\ Gain(y_i, \mathbf{X}_k) = H(y_i) - \frac{H(y_i)}{H(\mathbf{X}_k)}, \tag{14}$$

where, H denotes the entropy. Information gain has a shortcoming of giving favour to the features having large number of distinct values.

2) Gain Ratio: To overcome the aforementioned limitation of information gain, gain ratio is used that is presented as follows:

$$Gain\ Ratio(y_i, \mathbf{X}_k) = \frac{H(y_i) - \frac{H(y_i)}{H(\mathbf{X}_k)}}{H(y_1)}, \tag{15}$$

where, H denotes the entropy.

3) Chi-squared Statistic: Chi-squared statistic checks the independence of features values and class to which these values belong to. Null hypothesis states that the occurrence of these values are statistically independent. The greater value to statistic denotes the greater importance of feature against the null hypothesis. Chi-squared statistic is calculated by the following equation.

$$\chi^2\ statistic\ (y_i, \mathbf{X}_k, D) = \Sigma_{A_{\mathbf{X}_k}} \Sigma_{A_{y_i}} \frac{(O_{A_{\mathbf{X}_k} A_{y_i}} - E_{A_{\mathbf{X}_k} A_{y_i}})^2}{E_{A_{\mathbf{X}_k} A_{y_i}}}, \tag{16}$$

where, D denotes the dataset, $A_{\mathbf{X}_k}$ represents the attribute value of feature $\mathbf{X}_k$, A_{y_i} is the class value, $O_{A_{\mathbf{X}_k} A_{y_i}}$ refers to the observed value, and $E_{A_{\mathbf{X}_k} A_{y_i}}$ presents the expected value.

D. Performance Evaluation Metrics

Precision, Recall and balanced F-measure are standard metrics that are employed to check the performance of binary classification models. These state-of-the-art metrics are presented in Eq. (17, 18, 19). However, we mainly checked the performance of classification accuracy and RSP by using F-measure.

$$Precision = \frac{tp}{tp + tn} \tag{17}$$

$$Sensitivity = Recall = \frac{tp}{tp + fn} \tag{18}$$

$$F - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{19}$$

V. Conceptualizing Co-players

For a player, Co-player is the comrade who belongs to same or opponent team and have played matches during some common time span. A detailed description of a Co-player can be visualized from Figure 2. In our scenario, the concept of Co-player is taken as Co-batsmen and Co-bowlers.

To the best of our knowledge, we are the pioneers to put forward the concept of Co-players, Team and Opposite teams for RSP in sports domain. In details, a player has more chances of emerging as RS and subsequently becoming a domain star if it collaborates with the actual stars, and other rising stars during joint ventures. These collaborations provide an astonishing opportunity to player for learning the effective strategies and game plans from experienced domain experts and other comrades. Consequently, a player can uplift its abilities to maximum extent for becoming an actual star. Hence, the concept of Co-players is an important aspect to be considered for RSP.

A. Formulation of Features Space

This section describes the construction of features space based on graphs and contents information. The features are

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

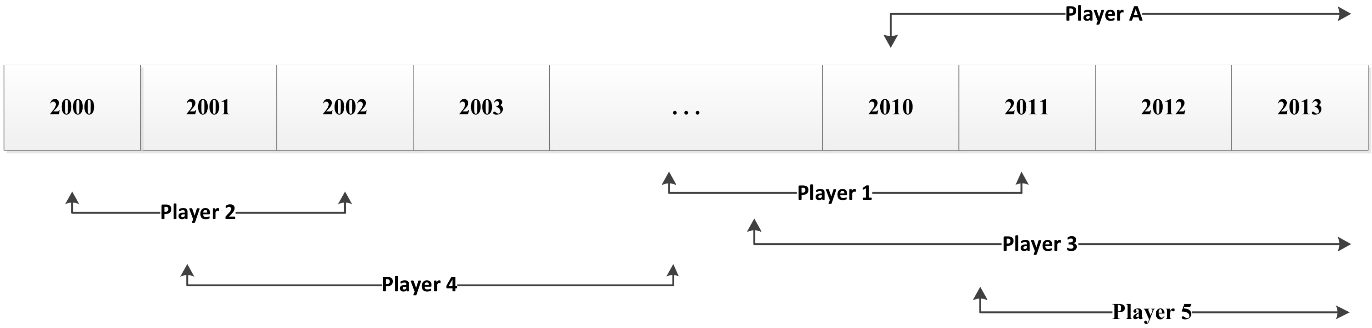


Fig. 2. Co-players of Player A: Players 1, 3 and 5 are Co-players of Player A as they share some common time span and might have joint ventures, while Player 2 and 4 are not Co-players of Player A.

defined for RSP in both domains, i.e. RSP in batting as well as in bowling realm. For each domain, three modules are considered and 9 features for batsmen, while 11 features for bowlers are taken into account as presented in Table I. Moreover, for Co-batsmen category, only those batsmen are considered who played from position 1 to 6. Similarly, while mining Co-bowlers for RSP, only bowlers from order 1 to 6 are considered. However, for Team and Opposite teams categories the scores of whole team players are included unlike former concepts. The mathematical notions of features for two major categories (Team and Opposite teams) are similar for both of the domains, but the dataset is tested and inquired from the respected domains. These features are utilized to train four classifiers. Subsequently, the performance of these trained classifiers is evaluated on unseen data, as presented in next section. For state-of-the-art definitions of metrics used in the features formulation, the readers are directed to subsection “Ranking metrics”. These metrics include runs, batsmen and bowlers (strike rate and average) and economy. Here, it is important to remark that we are extracting the features values only from those matches in which an emerging player (under investigation for RSP) has played i.e. joint ventures. In addition, all defined features have an effect on RSs performance, as these intend to motivate the RSs to learn better strategies from Co-players, host team and the opposite teams.

1) Co-batsmen Runs (CR): Runs made by the batsmen is the most salient feature considered for ranking purpose. CR of an emerging batsman (B) is the sum of all Co-batsmen runs ($\sum CoBR$). The mathematical notion of CR is presented as follows:

$$CR(B) = \sum CoBR. \tag{20}$$

2) Co-batsmen Average (CA): Average is the feature to check the consistent performance of a batsman. CA of a batsman (B) is the ratio of aggregated runs ($\sum CoBR$) to the total wickets fallen ($\sum CoBWF$) of all Co-batsmen presented as follows:

$$CA(B) = \frac{\sum CoBR}{\sum CoBWF}. \tag{21}$$

3) Co-batsmen Strike Rate (CSR): Strike rate is an effective measure to judge the batting performance of batsmen in limited overs game. CSR of a batsman (B) is the ratio of aggregate Co-batsmen runs ($\sum CoBR$) per 100 balls faced ($\sum CoBBF$) by all Co-batsmen as presented in the following equation:

$$CSR(B) = \frac{\sum CoBR}{\sum CoBBF}. \tag{22}$$

4) Team Average (TA): The feature of team average highly depends on the performance all the team players. Batsmen are supposed to make maximum runs while saving the wickets. Similarly, bowlers are expected to take more wickets while conceding minimum runs. TA (B) of a batsman (B) is defined to be the ratio of total runs scored ($\sum TR$) to total wickets fallen ($\sum TWF$) of a team. This feature is presented as follows:

$$TA(B) = \frac{\sum TR}{\sum TWF}, \tag{23}$$

whereas, team average of a particular emerging bowler TA (Bow) is the ratio of total runs conceded ($\sum TCR$) to the total wickets taken ($\sum TWT$) by the whole team bowlers. Mathematical notion is given as follows:

$$TA(Bow) = \frac{\sum TCR}{\sum TWT}. \tag{24}$$

5) Team Strike Rate (TSR): Strike rate is an important feature especially in ODI to check the performance of individuals or teams. For a batsman, it is the ratio of the runs scored to number of balls faced. Batsmen are supposed to score maximum runs in minimum balls. TSR (B) of an emerging batsman (B) is defined to be the ratio of total runs ($\sum TR$) made per 100 balls faced ($\sum BF$) by the team batsmen. The mathematical formulation of prescribed feature is presented as follows:

$$TSR(B) = \frac{\sum TR}{\sum BF}. \tag{25}$$

Team strike rate of a particular emerging bowler TSR (Bow) is the ratio of total balls bowled ($\sum TB$) to the total wickets taken ($\sum WT$) by all the team bowlers. Thus, the bowlers are

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

TABLE I
FEATURES DISTRIBUTION: THE COLUMN WITH TITLE CO-BATSMEN LISTS THE PRESCRIBED FEATURES OF BATSMEN, WHILE THE COLUMN ENTITLED WITH CO-BOWLERS PRESENTS THE BOWLERS FEATURES. THE MIDDLE COLUMNS (TEAM, OPPOSITE TEAMS) DENOTE THE DEFINED COMMON FEATURES FOR BOTH THE BATSMEN AND THE BOWLERS.

Batsmen Features	**Features for Batsmen/ Bowlers**		**Bowlers Features**
Co-batsmen	Team	Opposite teams	Co-bowlers
Co-batsmen Runs	Team Average (Batsman)	Opposite teams Average (Batsman)	Co-bowlers Economy
Co-batsmen Average	Team Strike Rate (Batsman)	Opposite teams Strike Rate (Batsman)	Co-bowlers Average
Co-batsmen Strike Rate	Team Win/Loss Ratio	Opposite teams Win/Loss Ratio	Co-bowlers Strike Rate
	Team Average (Bowler)	Opposite teams Average (Bowler)	
	Team Economy (Bowler)	Opposite teams Economy (Bowler)	
	Team Strike Rate (Bowler)	Opposite teams Strike Rate (Bowler)	

expected to take more wickets while within minimum balls. Mathematical notion is given as follows:

$$TSR(Bow) = \frac{\sum TB}{\sum WT}. \tag{26}$$

6) Team Economy (TE): Economy of a bowler is an influential feature considered for ranking purpose. TE of an emerging bowler (Bow) is the ratio of aggregate conceded runs ($\sum TCR$) to the total overs bowled ($\sum TO$) by team bowlers. The mathematical conception of TE (Bow) is presented as follows:

$$TE(Bow) = \frac{\sum TCR}{\sum TO}. \tag{27}$$

7) Team Win/Loss Ratio (T W/L): The ranking of a team highly depends on the win/loss ratio. The higher TW/L ratio depicts better batting and bowling performances by players against opposite teams. Formally,

$$TW/L = \frac{\sum TW}{\sum TL} \times \frac{TM - (NR + Tie)}{\sum TW + \sum TL}, \tag{28}$$

where, ($\sum TW$) and ($\sum TL$) denote the total number of win and loss of matches, respectively. The feature TM represents the total matches, NR points out the not resulted matches, while the term Tie is referring to the matches that ended as a draw i.e. no team could win that match. The factor ($\sum TL$) is removed from the denominator in case, if a team does not lose any match.

8) Opposite teams Average (OTA): OTA is an effective feature to check the batting and bowling performances of a team player against the all the opposite teams. OTA (B) of a particular batsman (B) is defined to be the average of ratio of total runs ($\sum OT_iR$) scored by the batsmen to total wickets fallen of opposite teams ($\sum OT_iWF$). This feature is presented as follows:

$$OTA(B) = \left\{ \sum_{i=1}^{n} \left(\frac{\sum OT_iR}{\sum OT_iWF} \right) \right\} /n, \tag{29}$$

while, opposite teams average of a particular emerging bowler OTA(Bow) is the average of ratio of total runs conceded by the team bowlers ($\sum OT_iCR$) to the total wickets taken ($\sum OT_iWT$) by them. Mathematical notion is given as follows:

$$OTA(Bow) = \left\{ \sum_{i=1}^{n} \left(\frac{\sum OT_iCR}{\sum OT_iWT} \right) \right\} /n. \tag{30}$$

9) Opposite teams Strike Rate (OTSR): OTSR is also considered as a supportive parameter to uplift the performance of an emerging player. OTSR of an emerging batsman (B) is defined to be the average of ratio of total runs ($\sum OT_iR$) made to total balls faced ($\sum OT_iBF$) by the opposite teams OT_i batsmen against the underlying team. The mathematical formulation of prescribed feature is presented as follows:

$$OTSR(B) = \left\{ \sum_{i=1}^{n} \left(\frac{\sum OT_iR}{\sum OT_iBF} \right) \right\} /n. \tag{31}$$

OTSR of an emerging bowler (Bow) is the average of ratio of total balls bowled ($\sum OT_iTB$) to the total wickets taken ($\sum OT_iWT$) by all opposite team bowlers. Thus, bowlers are expected to take more wickets within minimum balls. Mathematical notion is given as follows:

$$OTSR(Bow) = \left\{ \sum_{i=1}^{n} \left(\frac{\sum OT_iTB}{\sum OT_iWT} \right) \right\} /n. \tag{32}$$

10) Opposite teams Economy (OTE): OTE of an emerging bowler (Bow) is the ratio of aggregate conceded runs ($\sum OT_iCR$) to the total overs bowled ($\sum OT_iOB$) by team bowlers. The mathematical conception of OTE (Bow) is presented as follows:

$$OTE(Bow) = \left\{ \sum_{i=1}^{n} \left(\frac{\sum OT_iCR}{\sum OT_iOB} \right) \right\} /n. \tag{33}$$

11) Opposite teams Win/Loss Ratio (OT W/L): OT W/L is also an important feature that is considered while making a winning strategy against a particular opposite team OT_i. We

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

consider the factor OT W/L of all the opposite teams, in which an emerging player of a particular team has played. Formally,

$$OTW/L = \left\{ \sum_{i=1}^{n} \left(\frac{\sum OT_iW}{\sum OT_iL} \times \frac{TM_i - (NR_i + Tie_i)}{\sum OT_iW + \sum OT_iL} \right) \right\} /n, \tag{34}$$

where, ($\sum OT_iW$) and ($\sum OT_iL$) denote the total number of win and loss of matches by opposite teams OT_i against a particular team, respectively. The notion TM_i represents the total matches played by opposite teams OT_i against a particular team, NR_i points out the not resulted matches, while the term Tie_i is referring to the matches that ended as a draw i.e. no team could win that match. The factor ($\sum OT_iL$) is removed from the denominator in case, if a team does not lose any match.

12) Co-bowlers Average (CBA): CBA of a bowler (Bow) is the ratio of aggregate runs conceded ($\sum Co\,Bow\,CR$) to total wickets taken by all Co-bowlers ($\sum Co\,Bow\,WT$). Mathematically, this feature is presented as follows:

$$CBA(Bow) = \frac{\sum Co\,Bow\,CR}{\sum Co\,Bow\,WT}. \tag{35}$$

13) Co-bowlers Strike Rate (CBSR): CBSR of an e-merging bowler (Bow) is the ratio of total balls bowled ($\sum Co\,Bow\,TB$) to the total wickets taken ($\sum Co\,Bow\,WT$) by Co-bowlers, as presented in the following equation:

$$CBSR(Bow) = \frac{\sum Co\,Bow\,TB}{\sum Co\,Bow\,WT}. \tag{36}$$

14) Co-bowlers Economy (CBE): CBE of an emerging bowler (Bow) is the ratio of total conceded runs ($\sum Co\,Bow\,CR$) to the total overs bowled ($\sum Co\,Bow\,OB$) by Co-bowlers. Hence, CBE is an effective feature to be considered for RSP in bowling domain. The mathematical notion of CBE is presented as follows:

$$CBE(Bow) = \frac{\sum Co\,Bow\,CR}{\sum Co\,Bow\,OB}. \tag{37}$$

VI. Experiments and Evaluations

A. Dataset Acquisition

The data is taken from espncricinfo that is a reliable web forum containing data of all cricket matches ever played since 1779. We make RSP both for batsmen as well as for bowlers based on real world ODIs dataset ranging from 2006 to 2013. Moreover, the predictions are made for the players who started their ODI international career during the span 2006-2013, were having maximum age of 30 years until 2013, and are still playing in their respective international ODI teams. More precisely, the data for the span 2006 to 2013 is used for RSP of 2013 and onwards. But, the first four ODI years data of each RS candidate is incorporated for RSP. The reason for taking of such a long span is to incorporate the data of Co-players (especially to check the effects of domain stars on RSs performance). During underlying era, a total of 645 batsmen 560 bowlers have performed in 1138 ODI matches. However, a pre-processing step was performed to huge amount of data for extracting more relevant information for RSP. In details, the players who played less than 20 innings were eliminated from the dataset because they did not play matches with all top ranked teams.

After pre-processing, the batsmen were ranked in descending order w.r.t aggregate runs of individuals and bowlers are graded w.r.t. total wickets taken by individuals. Subsequently, top 200 instances for each domain (i.e. batsmen and bowlers) are picked, and their corresponding feature scores are extracted. Finally, two datasets for each domain are generated based on the following two metrics for RSP.

1) Weighted Average of Batsman (WA(B)): Runs (R), average (Avg) and strike rate (SR) are the three salient features that are considered to define the WA(B). Since all features are positively correlated with the batsman performance, hence, all of them are added. Besides, an equal threshold weightage of 33.33 is given to each factor as follows:

$$WA\,(B) = \frac{33.33 \times R + 33.33 \times Avg + 33.33 \times SR}{100}. \tag{38}$$

2) Weighted Average of Bowler (WA(Bow): Wickets (W), average (Avg), economy (Eco) and strike rate (SR) are the four salient features that are considered to define the WA (Bow). Since, number of wickets are positively correlated with the performance of bowlers, hence, this feature is added, while the remaining three features are subtracted due to the negative correlation with the bowlers performance. Besides, an equal threshold weight of 25 is given to each factor as follows:

$$WA\,(Bow) = \frac{25 \times W + 25 \times (-Avg) + 25 \times (-Eco) + 25 \times (-SR)}{100}. \tag{39}$$

The second dataset is extracted for year wise performance check based on different characteristic nominated as performance evolution of a batsman PE(B), which incorporates runs, average and strike rate of a batsman. In similar manner, performance evolution of a bowler PE (Bow) considers wickets, economy, average and strike rate of a bowler. First, we define the metric evolution that measures the ratio of change in the evolution indices (runs, average and strike rate of a batsman, while wickets, economy, average and strike rate of a bowler). Evolution of runs for a batsman i is defined as follows:

$$R\,evo_i = \frac{R_j - A_j}{A_j}, \tag{40}$$

where, R_j are the runs made by batsman i during year j and A_j is denoting the average runs made by all the comparative batsmen during the same year j. Similarly, $Avg\,evo_i$ and $SR\,evo_i$ capture the evolution of a batsman i while considering its average and strike rate, respectively. Likewise, $W\,evo_i$, $Eco\,evo_i$, $Avg\,evo_i$ and $SR\,evo_i$ are computed to capture the evolution of a bowler i.

3) Performance Evolution of a Batsman (PE (B)): Now, we formally present the mathematical notion for PE (B) that computes the year based evolution score of a batsman i as follows:

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

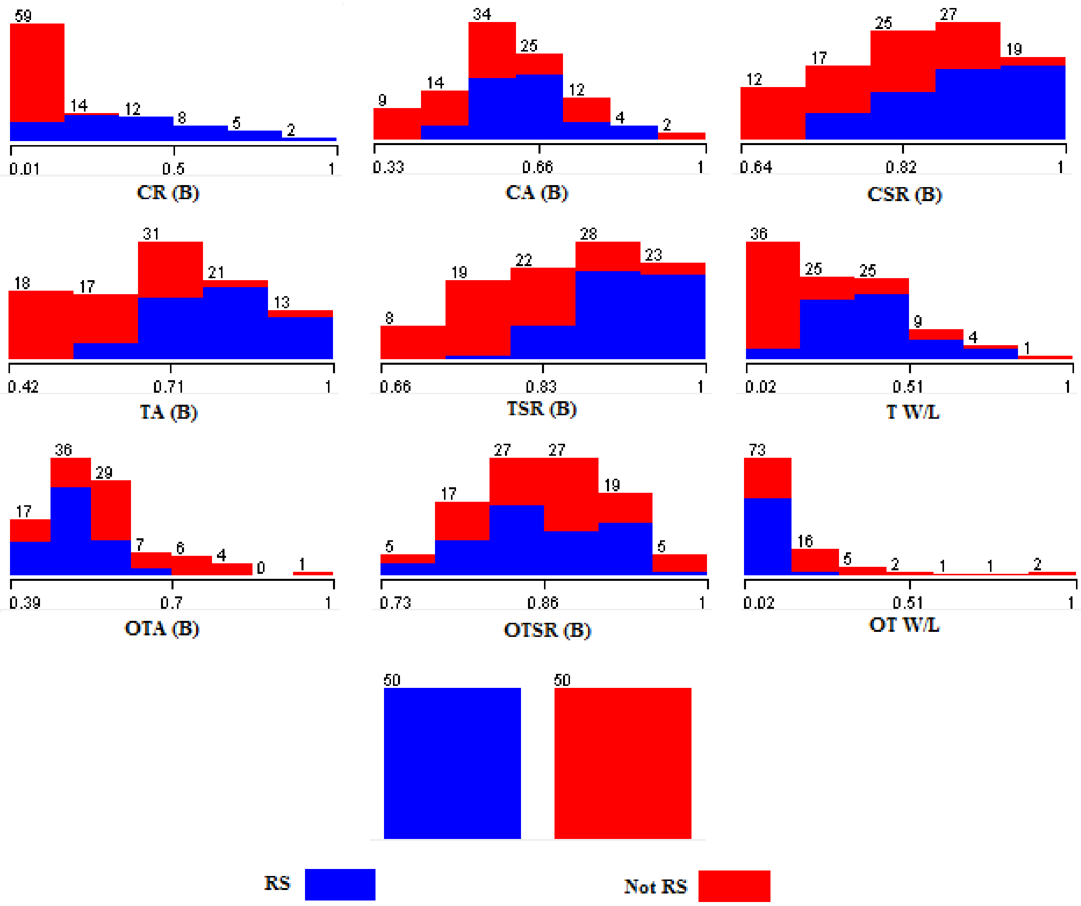


Fig. 3. Features statistical analysis using 1^{st} dataset (WA (B))

$$PE(B) = \frac{\sum_{i=1}^{4} R\ evo_i + \sum_{i=1}^{4} Avg\ evo_i + \sum_{i=1}^{4} SR\ evo_i}{4}, \tag{41}$$

where, 4 in denominator denotes first four years performance of an emerging batsman. The evolution indices appeared in numerator are added because all of them are positively correlated with the performance of a batsman.

4) Performance Evolution of a Bowler (PE (Bow)): Mathematical formulation of PE (Bow) is presented as follows:

$$PE(Bow) = \frac{\sum_{i=1}^{4} W\ evo_i - \sum_{i=1}^{4} Avg\ evo_i - \sum_{i=1}^{4} Eco\ evo_i - \sum_{i=1}^{4} SR\ evo_i}{4}, \tag{42}$$

where, 4 in denominator denotes first four years performance of an emerging bowler. Moreover, wickets taken by the bowler has positive correlation with its performance, therefore, it is added, while the remaining three features above the fraction have negative correlation with bowling performance, thus, that features values are subtracted.

Among 200 records belonging to each domain, 100 instances are representing players with the highest weighted average (a.k.a. RSs or positive samples), while other 100 instances are referring to players with the lowest weighted average (a.k.a. Not RSs or negative samples). Besides, 50 instances of positive samples and equally amount of negative samples are randomly chosen for training and testing of datasets. Same kind of measures are considered for extracting the second dataset. Hence, both of the datasets fulfill the requirement of balanced dataset because of comprising equal number of negative and positive labeled records.

B. Features Statistical Distribution

For the visualization of statistical distribution of the employed features among both classes, normalized datasets are plotted in the form of overlying bar graphs. Figure 3 is depicting the such overlying bar graphs that are presenting the statistical distributions of 9 features along with RS and Not RS classes of batting dataset (WA(B)). Since, Co-batsmen and Team categories are positively correlated with the performance of emerging batsmen as larger values indicate better performance, therefore, more features values belonging to these categories are closer to 1 for RS class. On contrary, features belonging to Opposite teams category are negatively correlated to the emerging batsmen performance, so, more values of these features are more closer to 1 for Not RS class. Figure 4 is illustrating the overlying bar graphs that are presenting the statistical distributions of 11 features along with RS and Not RS classes of bowling dataset (WA(Bow)). Since, the feature *T W/L* and Opposite team category (except *OT W/L*) are positively correlated with the performance of emerging bowlers as larger values indicate better performance,

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

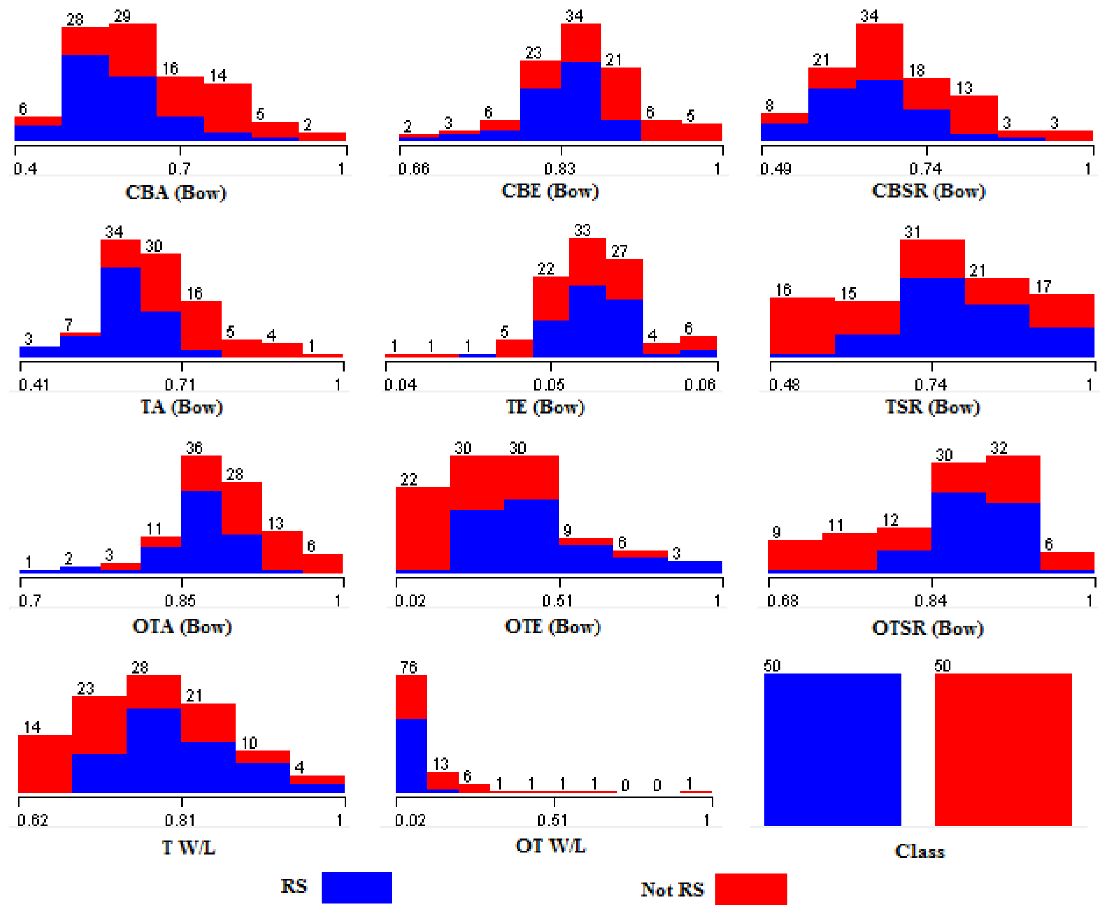


Fig. 4. Features statistical analysis using 1^{st} dataset (WA (Bow))

therefore, more values of these features are closer to 1 for RS class. On contrary, features belonging to Co-bowlers and Team categories are negatively correlated to the emerging bowlers performance, as higher values indicate poor performance, so, more features values of these categories are closer to 1 for Not RS class. Same distributions are observed with the remaining two datasets (PE (B) and (PE (Bow)), but are not presented due to conciseness.

C. Features Evaluation

This subsection presents the details about the relevance and importance of defined features for RSP. More precisely, information gain, gain ratio and chi-squared statistic are employed in order to check the relative importance of all underlying features for emerging batsmen and bowlers. After applying these state-of-the-art metrics, the features are ranked with respect to resulting values. The ranks of batting features by employing both datasets are depicted in Table II, while that of bowlers features ranking is presented in Table III. Although, the features from both domains are getting different rank values, but all of the features have the information participation for RSP as none among them has got the zero numerical value.

D. Performance Evaluation

We employed 10-fold cross validation procedure in order to train and validate the classifiers while using both datasets for each domain. For analyzing the effect of each feature on binary classification for RSP, each classifier is trained while exploiting each feature. This process is accomplished in total of 9 learning cycles for batsmen, while 11 for bowling domain. Each dataset comprising of 100 labeled samples is segregated into 10 equal parts, such as $10, 20, \ldots, 100$. Subsequently, for each partition, all classifiers are trained. The overall learning setup is accomplished by employing an open source software WEKA. Finally, precision, recall and F-measure are computed for analyzing the results. However, we only present the results of average F-measure for analysis purpose. Based on RSs feature scores, RSP is made for both domains and presented in the subsequent section. For checking the effectiveness of selected features, we cross checked the ranking presented by us with that of provided by ICC from 2013-16. All evaluations reported in this paper are presenting the average of 10 observations of F-measure. Thus, average F-measure is plotted for each feature of both domains and their corresponding results are depicted in Figures 5-16.

E. Individual Feature Analysis

1) Batting Domain: Individual features are examined by employing state-of-the-art machine learning algorithms. The effectiveness of these distinct batting features for RSP is analyzed and presented in this subsection. For first dataset based on WA (B) that undertakes overall performance of a

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

TABLE II
BATTING FEATURES RANKING: RANKING COMPARISON OF BATSMEN FEATURES EXTRACTED FROM DEFINED DATASETS

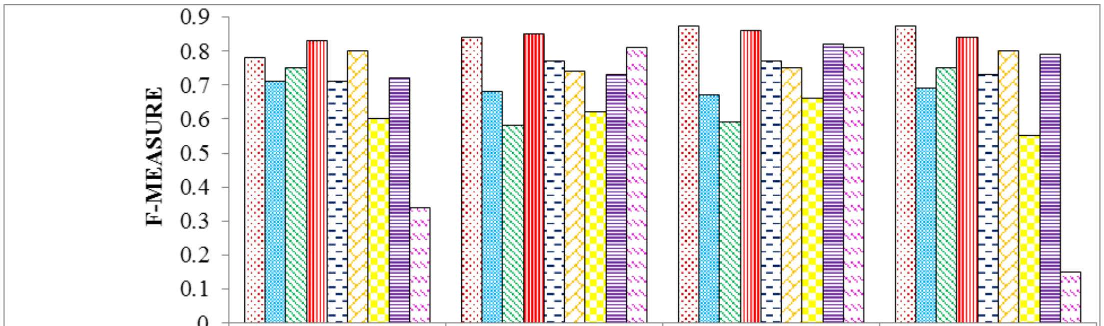
Rank	1^{st} dataset (WA (B))						2^{nd} dataset (PE (B))					
	Attribute Name	Info Gain	Attribute Name	Gain Ratio	Attribute Name	Chi-Squared Statistic	Attribute Name	Info Gain	Attribute Name	Gain Ratio	Attribute Name	Chi-Squared Statistic
1	CR (B)	0.512	CR (B)	0.558	CR (B)	62.87	CR (B)	0.512	CR (B)	0.516	CR (B)	61.45
2	TSR (B)	0.393	TA (B)	0.416	TA (B)	51.84	TSR (B)	0.393	TSR (B)	0.394	TSR (B)	49.17
3	TA (B)	0.383	T W/L	0.369	T W/L	43.46	TA (B)	0.383	TA (B)	0.39	TA (B)	47.45
4	T W/L	0.317	TSR (B)	0.321	TSR (B)	41.02	T W/L	0.317	T W/L	0.336	T W/L	39.06
5	OT W/L	0.245	OT W/L	0.305	OT W/L	37.5	OT W/L	0.245	CSR (B)	0.263	OT W/L	31.27
6	CSR (B)	0.235	CSR (B)	0.231	CSR (B)	34.91	CSR (B)	0.235	OT W/L	0.258	CSR (B)	29.21
7	OTA (B)	0.179	OTA (B)	0.209	OTA (B)	26.27	OTA (B)	0.179	OTA (B)	0.191	OTA (B)	23.25
8	CA (B)	0.131	CA (B)	0.158	CA (B)	19.38	CA (B)	0.131	CA (B)	0.158	CA (B)	16.83
9	OTSR (B)	0.10	OTSR (B)	0.11	OTSR (B)	10.08	OTSR (B)	0.10	OTSR (B)	0.11	OTSR (B)	9.8

TABLE III
BOWLING FEATURES RANKING: RANKING COMPARISON OF BOWLING FEATURES EXTRACTED FROM DEFINED DATASETS

Rank	1^{st} dataset (WA (Bow))						2^{nd} dataset (PE (Bow))					
	Attribute Name	Info Gain	Attribute Name	Gain Ratio	Attribute Name	Chi-Squared Statistic	Attribute Name	Info Gain	Attribute Name	Gain Ratio	Attribute Name	Chi-Squared Statistic
1	TA (Bow)	0.284	TA (Bow)	0.287	TA (Bow)	36.52	TSR (Bow)	0.235	TA (Bow)	0.297	OTA (Bow)	29.21
2	T W/L	0.212	OT W/L	0.273	TE (Bow)	25	OTA (Bow)	0.235	TSR (Bow)	0.275	TSR (Bow)	28.57
3	OT W/L	0.212	T W/L	0.273	OT W/L	24.90	T W/L	0.212	T W/L	0.273	OT W/L	24.90
4	TSR (Bow)	0.194	OTSR (Bow)	0.268	T W/L	24.90	OTE (Bow)	0.212	OT W/L	0.273	T W/L	24.90
5	TE (Bow)	0.193	CBE (Bow)	0.258	TSR (Bow)	27.73	OT W/L	0.212	OTE (Bow)	0.273	OTE (Bow)	24.90
6	OTA (Bow)	0.166	TSR (Bow)	0.218	OTA (Bow)	20.38	TA (Bow)	0.195	OTA (Bow)	0.263	OTSR (Bow)	21.56
7	OTSR (Bow)	0.157	OTE (Bow)	0.216	CBA (Bow)	17.82	OTSR (Bow)	0.165	OTSR (Bow)	0.179	TA (Bow)	20.48
8	OTE (Bow)	0.147	OTA (Bow)	0.214	CBSR (Bow)	17.76	CBA (Bow)	0.131	CBA (Bow)	0.13	CBA (Bow)	16.83
9	CBE (Bow)	0.144	TE (Bow)	0.204	OTE (Bow)	17.34	TE (Bow)	0.11	TE (Bow)	0.11	TE (Bow)	13.8
10	CBSR (Bow)	0.141	CBSR (Bow)	0.177	OTSR (Bow)	16.27	CBSR (Bow)	0.101	CBSR (Bow)	0.102	CBSR (Bow)	11.83
11	CBA (Bow)	0.135	CBA (Bow)	0.146	CBE (Bow)	14.94	CBE (Bow)	0.10	CBE (Bow)	0.1	CBE (Bow)	10.83

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access



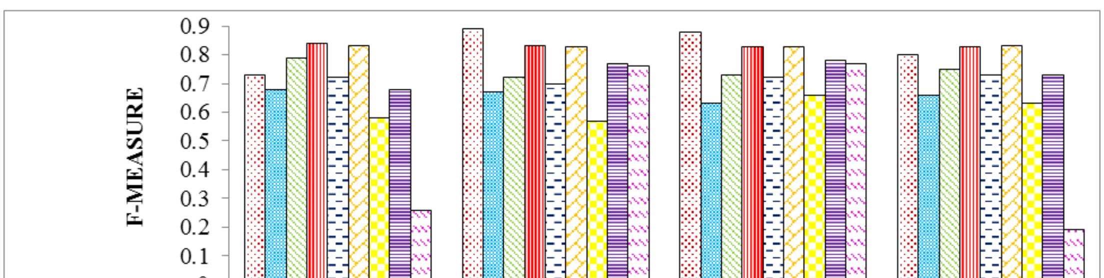
	SVM	CART	BN	NB
CR (B)	0.78	0.84	0.875	0.873
CA (B)	0.71	0.68	0.67	0.69
CSR (B)	0.75	0.58	0.59	0.75
TA (B)	0.83	0.85	0.86	0.84
OTA (B)	0.71	0.77	0.77	0.73
TSR (B)	0.8	0.74	0.75	0.8
OTSR (B)	0.6	0.62	0.66	0.55
T W/L	0.72	0.73	0.82	0.79
OT W/L	0.34	0.81	0.81	0.15

Fig. 5. Features F-measure analysis using 1^{st} dataset (WA (B))

batsman, CR (B) dominates all other features for RSP with the accuracy of 87.5%. Therefore, we remark that only CR (B) can predict RSs more effectively as compared to the other features. More precisely, for the same feature, we get the accuracies of 87.5%, 87.3%, 84% and 78% by applying BN, NB, CART and SVM, respectively.

Moreover, we find TA (B) and TSR (B) as second and third best feature for RSP. Overall, BN classifier surpasses all the remaining classifiers for RSP, while the precedence order of remaining classifiers in terms of F-measure accuracy for first dataset is found as CART, SVM and NB, respectively. While taking into account WA (B) based dataset results presented in Figure 5, we deduce that all defined features are effective for RSP in batting domain.

The same experiment is performed on the second batting dataset that is based on the year wise performance evolution measure nominated as PE (B). For this dataset, CR (B) suppresses all the remaining features while attaining the accuracy of 89 %. This outcome also fortifies our claim that individually CR (B) is the best feature among the others for RSP. In more

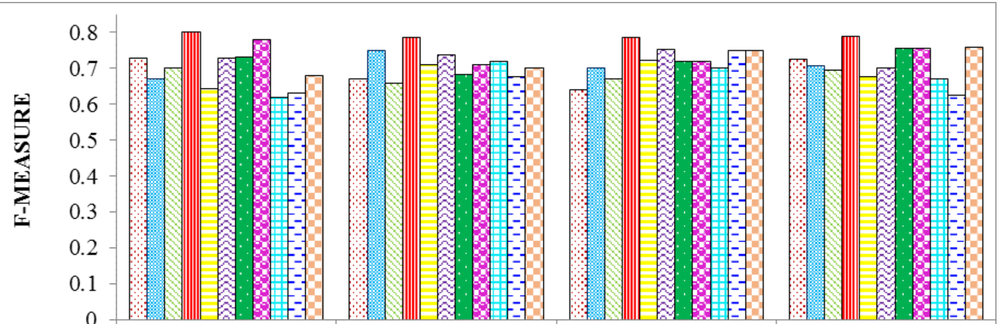


	SVM	CART	BN	NB
CR (B)	0.73	0.89	0.88	0.8
CA (B)	0.68	0.67	0.63	0.66
CSR (B)	0.79	0.72	0.73	0.75
TA (B)	0.84	0.83	0.826	0.827
OTA (B)	0.72	0.7	0.72	0.73
TSR (B)	0.83	0.827	0.827	0.83
OTSR (B)	0.58	0.57	0.66	0.63
T W/L	0.68	0.77	0.78	0.73
OT W/L	0.26	0.76	0.77	0.19

Fig. 6. Features F-measure analysis using 2^{nd} dataset (PE (B))

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

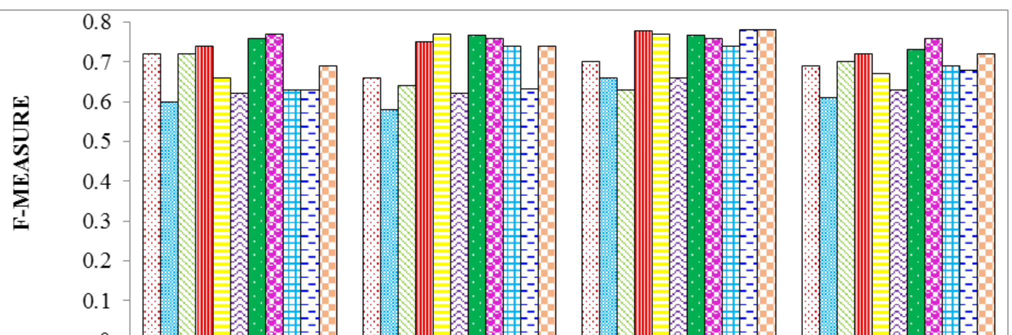
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access



	SVM	CART	BN	NB
CBA (Bow)	0.729	0.67	0.64	0.724
CBE (Bow)	0.67	0.75	0.702	0.708
CBSR (Bow)	0.7	0.66	0.672	0.695
TA (Bow)	0.8	0.785	0.785	0.788
OTA (Bow)	0.642	0.71	0.723	0.678
TE (Bow)	0.729	0.737	0.752	0.7
OTE (Bow)	0.732	0.683	0.718	0.756
TSR (Bow)	0.78	0.71	0.72	0.757
OTSR (Bow)	0.62	0.72	0.7	0.67
T W/L	0.63	0.678	0.75	0.624
OT W/L	0.68	0.7	0.75	0.76

Fig. 7. Features F-measure analysis using 1^{st} dataset (WA (Bow))

details, for the same feature, we get 89%, 88%, 80% and 73% accuracies while employing CART, BN, NB and SVM classifiers, respectively. Moreover, TA (B) and TSR (B) are found as second and third best features for the prediction of RSs. Furthermore, OT W/L is the feature that stands at last in terms of classification accuracy, thus, it is deduced that this feature effects the least to improve the performance of RS. Overall, BN classifier dominates all the remaining classifiers for RSP, while the precedence order of remaining classifiers in terms of F-measure accuracy for second dataset is found as CART, NB and SVM, respectively. While taking into account PE (B) based dataset results presented in Figure 6,



	SVM	CART	BN	NB
CBA (Bow)	0.72	0.66	0.7	0.69
CBE (Bow)	0.6	0.58	0.66	0.61
CBSR (Bow)	0.72	0.64	0.63	0.7
TA (Bow)	0.74	0.75	0.778	0.72
OTA (Bow)	0.66	0.77	0.77	0.67
TE (Bow)	0.62	0.62	0.66	0.63
OTE (Bow)	0.76	0.766	0.766	0.73
TSR (Bow)	0.77	0.76	0.76	0.76
OTSR (Bow)	0.63	0.739	0.739	0.69
T W/L	0.63	0.632	0.78	0.68
OT W/L	0.69	0.74	0.78	0.72

Fig. 8. Features F-measure analysis using 2^{nd} dataset (PE (Bow))

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

we conclude that all defined features are effective for RSP in batting domain. Additionally, it is observed that the influence of individual features is same in terms of performance and classification accuracy for both datasets. This concludes the brief analysis of distinct features impact on RSP for batting domain.

2) Bowling Domain: Individual bowling features are examined by employing state-of-the-art machine learning algorithms. The effectiveness of these distinct bowling features for RSP is analyzed and presented in this subsection. For first dataset based on WA (Bow) that undertakes overall performance of a bowler, TA (Bow) dominates all other features for RSP with the accuracy of 80%. Therefore, we remark that only TA (Bow) can predict RSs more effectively as compared to other features. More precisely, for the same feature, we get the accuracies of 80%, 78.8%, 78.5% and 78.5% by applying SVM, NB, CART, and BN, respectively. Moreover, we find TSR (Bow) and TE (Bow) as second and third best feature for RSP. Overall results demonstrate that BN classifier surpasses all remaining classifiers for RSP, while the precedence order of remaining classifiers in terms of F-measure accuracy for first dataset is found as NB, CART and SVM, respectively. While taking into account WA (Bow) based dataset results presented in Figure 7, we deduce that all defined features are effective for RSP in bowling domain.

Similar evaluation is performed on the second bowling dataset that is based on the year wise performance evolution measure nominated as PE (Bow). For this dataset, TA (Bow) suppresses all the remaining features while attaining the accuracy of 77.8%. This outcome also fortifies our claim that individually TA (Bow) is the best feature among others for RSP in bowling domain. In more details, for the same feature, we get 77.8%, 75%, 74% and 72% accuracies while employing BN, CART, SVM and NB classifiers, respectively. Moreover, T W/L and TSR (Bow) are found as second and third best features for the prediction of RSs. Overall, BN classifier dominates all the remaining classifiers for RSP, while the precedence order of remaining classifiers in terms of F-measure accuracy for second dataset is found as CART, NB and SVM, respectively. While taking into account PE (Bow) based dataset results presented in Figure 8, we conclude that all defined features are effective for RSP in bowling domain. Additionally, it is observed that the influence of individual features is bit different in terms of performance and classification accuracy because the datasets are extracted based on different statistical measures. This concludes the brief analysis of distinct features impact on RSP in bowling domain.

F. Category Wise Analysis

1) Batting Domain: This subsection presents the detailed analysis of category wise performance of defined features for batting domain. The features CR (B), CA (B) and CSR (B) are included in Co-batsmen category. Team category is comprised of TA (B), TSR (B) and T W/L, while the features OTA (B), OTSR (B) and OT W/L are counted in Opposite teams category. First dataset that is extracted based on overall performance metric WA (B) for RSP in batting domain shows

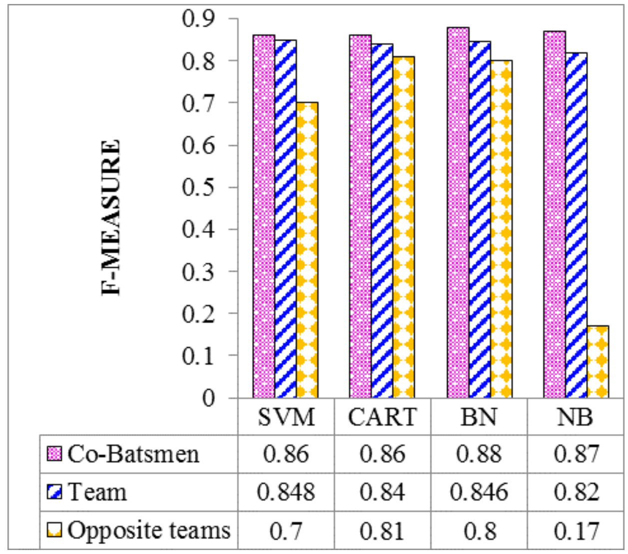


Fig. 9. 1^{st} dataset (WA (B))

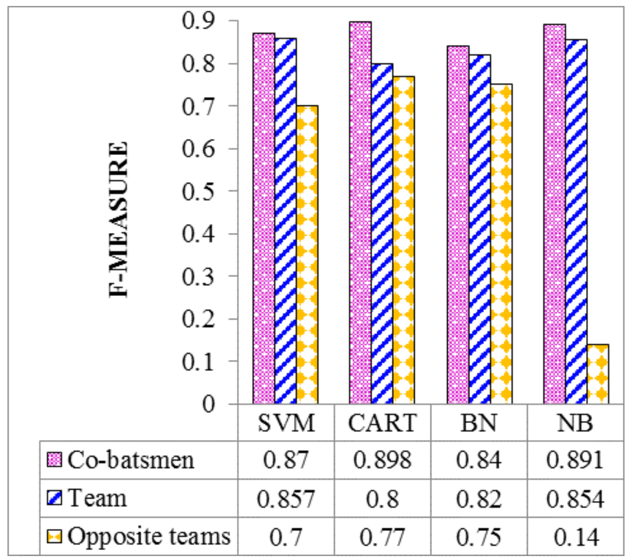


Fig. 10. 2^{nd} dataset (PE (B))

outclass performance as presented in Figure 9. The Co-batsmen category outperforms the other two categories while attaining 88 % F-measure accuracy. Thus, we infer that Co-batsmen category is better for RSP as compared to remaining two categories. In more details, for the same category, we get 88%, 87%, 86% and 87% learning accuracies by applying BN, NB, CART and SVM, respectively. Moreover, we find Team as second best category, while Opposite teams category acquires the lowest accuracy scores. Overall, BN dominates all other classifiers in terms of learning for RSP. The precedence order of remaining classifiers in terms of F-measure accuracy for first dataset is found as CART, SVM and NB, respectively.

Same evaluations are performed on the second batting dataset that is based on the year wise performance evolution measure nominated as PE (B), and the results are depicted in Figure 10. For this dataset also, Co-batsmen category

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

suppresses the remaining two categories while attaining the accuracy of 89.8 %. This outcome also fortifies our claim that individually Co-batsmen is the best category among the others for RSP. In more details, for the same category, we get 89.8%, 89.1%, 87% and 84% accuracies while employing CART, NB, SVM and BN classifiers, respectively. Moreover, we find Team as second best category, while Opposite teams category acquires the lowest accuracy scores. Overall, CART classifier dominates all the remaining classifiers for RSP, while the precedence order of remaining classifiers in terms of F-measure accuracy for second dataset is found as SVM, BN and NB, respectively. Additionally, it is observed that the influence of individual categories is same in terms of performance and classification accuracy for both datasets. This completes the brief category wise analysis of RSP in batting domain.

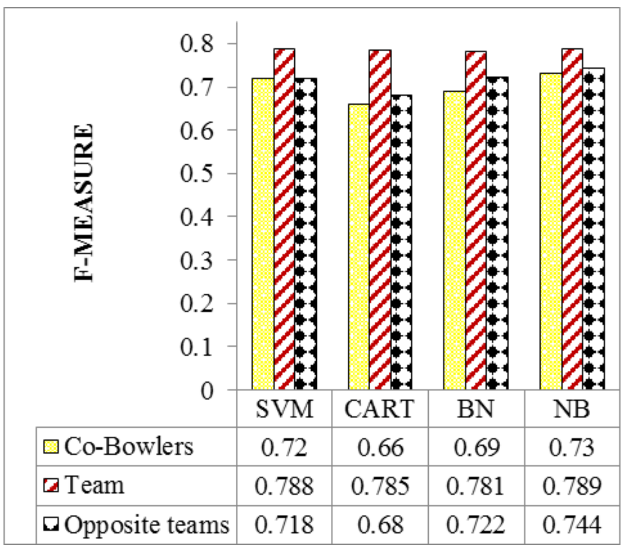


Fig. 11. 1^{st} dataset (WA (Bow))

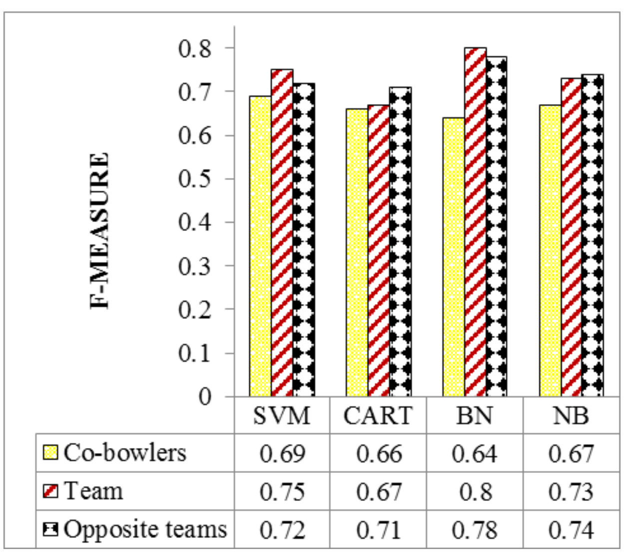


Fig. 12. 2^{nd} dataset (PE (Bow))

2) Bowling Domain: This subsection presents the detailed analysis of categories wise performance of defined features for bowling domain. The features CBA (Bow), CBE (Bow) and CBSR (Bow) are included in Co-bowlers category. Team category is comprised of TA (Bow), TE (Bow), TSR (Bow) and T W/L, while the OTA (Bow), OTE (Bow), OTSR (Bow) and OT W/L are counted in Opposite teams category.

First dataset that is extracted based on overall performance metric WA (Bow) for RSP in bowling domain shows better performance as presented in Figure 11. Unlike batting domain, the Team category outperforms the other two categories while attaining 78.9 % F-measure accuracy. Thus, we infer that Team category is better for RSP as compared to remaining two categories. In more details, for the same category, we get 78.9%, 78.8%, 78.5% and 78.1% learning accuracies by applying NB, SVM, CART and NB, respectively. Moreover, we find Opposite teams as second best category, while Co-bowlers category acquires the lowest accuracy scores. Overall, NB dominates all other classifiers in terms of learning for RSP. The precedence order of remaining classifiers in terms of F-measure accuracy for first dataset is found as SVM, BN and CART, respectively.

Same evaluations are performed on the second bowling dataset that is based on the year wise performance evolution measure nominated as PE (Bow), and the results are depicted in Figure 12. For this dataset also, Team category suppresses the remaining two categories while attaining the accuracy of 80 %. This outcome also fortifies that Team is the best individual category among the others for RSP. In more details, for the same category, we get 80%, 75%, 73% and 67% accuracies while employing BN, SVM, NB and CART classifiers, respectively. Moreover, we find Opposite teams as second best category, while Co-bowlers category acquires the lowest accuracy scores. Overall, BN classifier dominates all the remaining classifiers for RSP, while the precedence order of remaining classifiers in terms of F-measure accuracy for second dataset is found as SVM, NB and CART, respectively. Additionally, it is observed that the influence of individual categories is same in terms of performance and classification accuracy for both datasets. This concludes the brief category wise analysis of RSP for bowling domain.

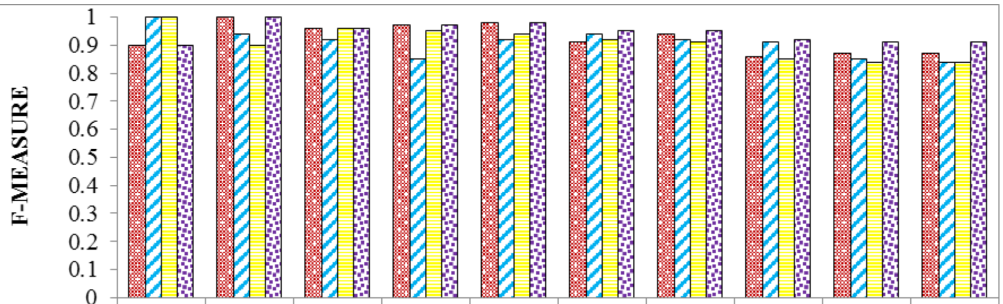
G. Model Wise Analysis

The comparative analysis of average F-measure accuracy results by applying generative and discriminative models for RSP in both domains are presented in this subsection.

1) Batting Domain: We analyze the impact of defined features for RSP and find that all state-of-the-art classifiers are showing outclass performance. The underlying subsection provides the analysis for learning RSs from the defined features for WA (B) measure based batting dataset. Thus, proposed features can be generalized for RSP in cricket domain. For different number of instances, every classifier is predicting RSs with 100% accuracy. However, overall NB is dominating all the remaining classifiers while achieving the average of 94.5% learning accuracy for 10-100 instances. The second best performance is exposed by SVM model with the average of

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access



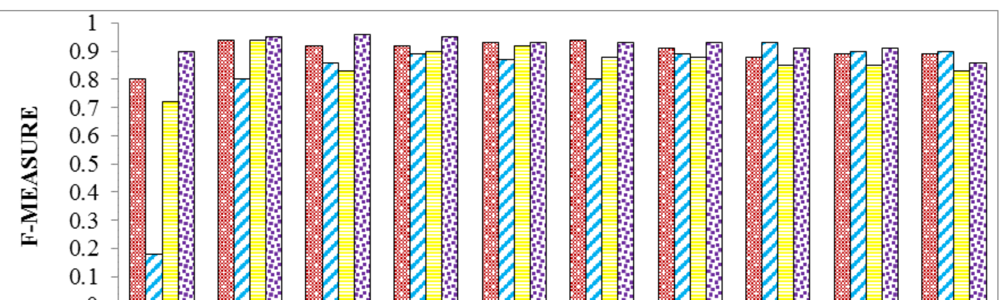
	10	20	30	40	50	60	70	80	90	100
SVM	0.9	1	0.96	0.97	0.98	0.91	0.94	0.86	0.87	0.87
CART	1	0.94	0.92	0.85	0.92	0.94	0.92	0.91	0.85	0.84
BN	1	0.9	0.96	0.95	0.94	0.92	0.91	0.85	0.84	0.84
NB	0.9	1	0.96	0.97	0.98	0.95	0.95	0.92	0.91	0.91

Fig. 13. Model wise analysis of features using 1^{st} dataset (WA (B)).

92.6% accuracy. BN stands at third with the average of 91.1% learning accuracy, while CART is ranked at last with the average of 90.1% accuracy for 10-100 instances. We observe from the Figure 13 that the accuracy is somehow decreasing a bit with the increase in number of instances. This bitsy decrement is due to inclusion of Opposite teams category, however, if we exclude this category, we get almost consistent results for underlying dataset. In general for this dataset, as a group, generative models are providing better results as compared to discriminative models.

Same experiment is performed on the second batting dataset that is based on the year wise performance evolution measure nominated as PE (B) and results are presented in Figure 14. For this dataset, NB suppresses all the remaining models while attaining the maximum accuracy of 96%. In more details, for all features, we get maximum 96%, 94%, 94% and 93% accuracies while employing NB, BN, SVM and CART classifiers, respectively. Overall, NB is dominating all the remaining classifiers while attaining the average accuracy of 92.3% for 10-100 instances. The precedence order of remaining classifiers for 10-100 instances of second dataset is found as SVM, BN and CART with the average F-measure accuracies of 90.2%, 86% and 80.2%, respectively. It is observed that the influence of individual features is consistent in terms of performance and classification accuracy, however, it is a bit lesser than the accuracy of first dataset due to the usage of different statistical metric for evaluation. In general, also for this dataset, as a group, generative models are providing better results as compared to discriminative models. Thus, we conclude that all defined features are effective for RSP in batting domain. This completes the brief analysis of distinct features impact on RSP in batting domain.

2) Bowling Domain: We examine the influence of defined features for RSP and find that all the state-of-the-art classifiers are showing excellent performance. The underlying subsection provides the analysis for learning RSs from the defined features for WA (Bow) measure based bowling dataset. Thus, proposed features can be generalized for RSP in cricket domain. For this dataset, NB suppresses all the remaining models while attaining the maximum accuracy of 87%. In more details, for all features, we get maximum 87%, 84%, 82.4% and 80% accuracies while employing NB, SVM, B-



	10	20	30	40	50	60	70	80	90	100
SVM	0.8	0.94	0.92	0.92	0.93	0.94	0.91	0.88	0.89	0.89
CART	0.18	0.8	0.86	0.89	0.87	0.8	0.89	0.93	0.9	0.9
BN	0.72	0.94	0.83	0.9	0.92	0.88	0.88	0.85	0.85	0.83
NB	0.9	0.95	0.96	0.95	0.93	0.93	0.93	0.91	0.91	0.86

Fig. 14. Model wise analysis of features using 2^{nd} dataset (PE (B)).

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

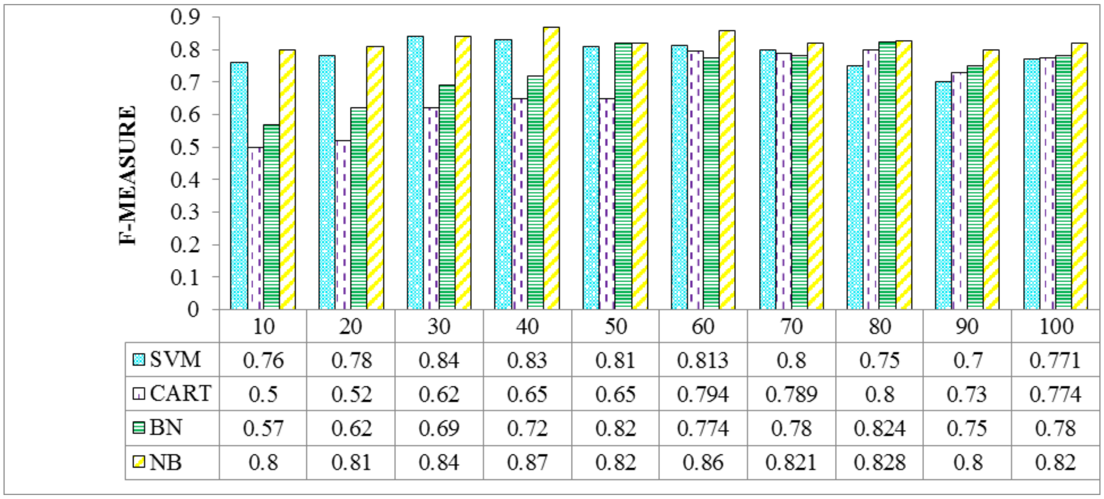


Fig. 15. Model wise analysis of features using 1^{st} dataset (WA (Bow)).

N and CART classifiers, respectively. However, overall NB demonstrates the best performance while achieving the average of 82.69% learning accuracy for 10-100 instances. The second best performance is exposed by SVM model with the average of 78.54% accuracy. BN stands at third with the average of 73.28% learning accuracy, while CART is ranked at last with the average of 68.27% accuracy for 10-100 instances. In general for this dataset, as a group, generative models are providing better results as compared to discriminative models. We can observe from the Figure 15 that the accuracy is almost consistent with the increase in number of instances.

Similar experiment is performed on the second bowling dataset that is based on the year wise performance evolution measure nominated as PE (Bow) and results are presented in Figure 16. For this dataset, SVM suppresses all the remaining models while attaining the maximum accuracy of 87%. In more details, for all features, we get maximum 87%, 85%, 79% and 76% accuracies while employing SVM, NB, BN and CART classifiers, respectively. However, SVM is dominating all the remaining classifiers while attaining the average accuracy of 81.8% for 10-100 instances. The precedence order of remaining classifiers for 10-100 instances of second dataset is found as NB, BN and CART with the average F-measure accuracies of 79.5%, 73.1% and 65.6%, respectively. It is observed that the influence of individual features is consistent in terms of performance and classification accuracy, however, it is a bit lesser than the accuracy of first dataset due to the usage of different statistical metric for evaluation. In general, for this dataset also, as a group, generative models are providing better results as compared to discriminative models. Overall, for the underlying datasets, generative models demonstrate better performance as compared to discriminative models. Finally, we conclude that all defined features are effective for RSP in bowling domain. This concludes the brief model wise analysis of distinct features impact on RSP for bowling domain.

VII. Rankings of RSs

A. Batting Domain

For batting domain, we present the ranking of top 10 RS batsmen based on WA (B), PE (B) and RS (B). The metrics WA (B) and PE (B) are formally presented in the previous subsection. The third metric RS (B) is composed

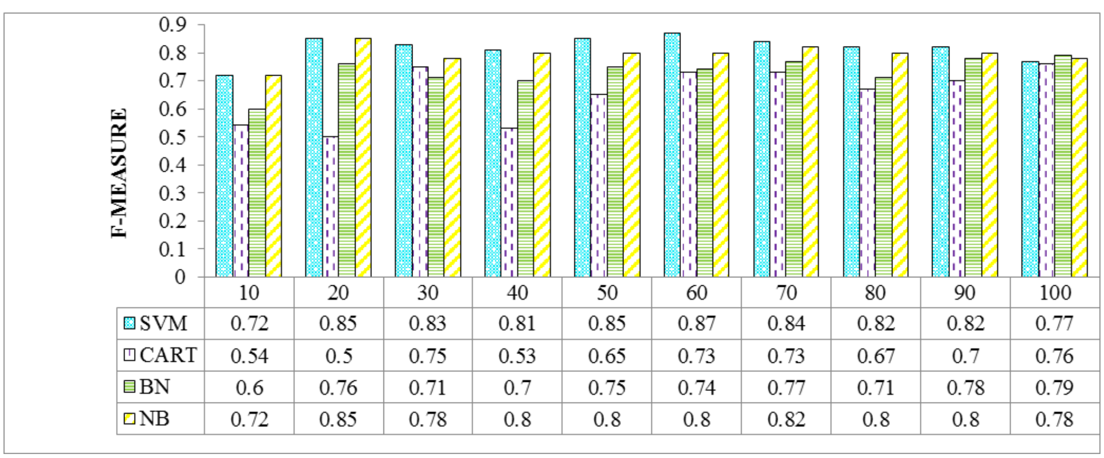


Fig. 16. Model wise analysis of features using 2nd dataset (PE (Bow)).

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

of aggregate score that is calculated by adding all the positively correlated features to the batting performance, while the negative correlated features are subtracted. More precisely, among the 9 defined features for RS, 6 features belonging to Co-batsmen and Team categories are positively correlated to the RS batsmen performance, because higher values of these features indicate the higher chance for an emerging batsman of becoming a RS. On contrary, the three features of Opposite teams category are negatively correlated with the performance of batsman. Formally, the RS score for a batsman (B) is calculated as follows:

$$\begin{aligned} RS(B) = CR(B) + CA(B) + CSR(B) \\ +TA(B) + TSR(B) + TW/L \\ -OTA(B) - OTSR(B) - OTW/L. \end{aligned} \tag{43}$$

Here, we mention that all distinct features have different ranges with respect to scores, therefore, to incorporate equal weightage of each feature, we first normalize the feature values in the range of 0 to 1 and then calculate the scores for all metrics. Subsequently, the RS batsmen are organized in descending order and leading 10 are listed in Table IV. Each ranking based on aforementioned three metrics in Table IV is depicting the detail of player name, the country it belongs to and the highest ranking it acquired by ICC from 2013-16. It is clear from the Table IV that predicted top ten batting RSs are ranked among top 13, 13 and 23 w.r.t. WA (B), PE (B) and RS (B) score, respectively.

Besides, we comment that these RSs have got such astonishing positions while competing with the star players of batting domain. Among top ten RSs based on WA (B) and PE (B), there are 7 communal batsmen. Similarly, rankings based on WA (B) and RS (B) present 6 in common, while that of PE (B) and RS (B) based rankings depict 7 communal among top ten predicted RSs. However, ranking varies for each RS batsman w.r.t. different metrics due to the incorporation of different statistical measures. For instance, AJ Finch has got a maximum of 7^{th} rank by ICC in 2015 and he is on 1^{st} position, while V. Kohli acquired 1^{st} rank by ICC and he is on 2^{nd} number as per ranked by RS (B) score. The reason is that AJ Finch has played less matches, therefore, he has much better CA (B), CSR (B) and T W/L scores as compared to V. Kohli. These features show a bitsy declining trend as more Co-batsmen are involved or more matches are played by underlying player. Similar causes are observed with the case of AM Rahane but based on our detailed analysis, we predict that in the near future, he will get much higher rank by ICC as compared to his current rank. Nonetheless, all the presented batsmen are RSs for sure as per validated by ICC rankings.

B. Bowling Domain

For bowling domain, we present the ranking of top 10 RS bowlers based on WA (Bow), PE (Bow) and RS (Bow). The metrics WA (Bow) and PE (Bow) are formally presented in the former subsection. The third metric RS (Bow) is composed of aggregate score that is calculated by adding all the positively correlated features to the bowling performance, while the negative correlated features are subtracted. More precisely, among the 11 defined features for bowling domain, three features of Opposite teams category along with the feature T W/L are positively correlated to the RS bowlers performance, because higher values of these features indicate the higher chance for an emerging bowler of becoming a RS. On contrary, 7 features belonging to Co-bowlers and Team categories (except T W/L) are negatively correlated with the performance of bowler. Formally, the RS score for a bowler (Bow) is calculated as follows:

$$\begin{aligned} RS(Bow) = TW/L + OTE(Bow) \\ +OTA(Bow) + OTSR(Bow) - CBA(Bow) \\ -CBE(Bow) - CBSR(Bow) - TA(Bow) \\ -TE(Bow) - TSR(Bow) - OTW/L \end{aligned} \tag{44}$$

We first normalize the feature values in the range of 0 to 1 and then calculate the scores for all metrics. Subsequently, the RS bowlers are organized in descending order and the leading 10 are listed in Table V. Each ranking based on aforementioned three metrics is depicting the detail of player name, the country it belongs to and the highest ranking it acquired by ICC from 2013-16. It is clear from the Table V that predicted leading ten bowling RSs are ranked among top 19, 09 and 21 w.r.t. WA (Bow), PE (Bow) and RS (Bow) score, respectively. Further, we comment that these RSs have got such astonishing positions while competing with the star players of bowling domain. Among top ten RSs based on WA (Bow) and PE (Bow), there are 9 communal bowlers. Similarly, rankings based on WA (Bow) and RS (Bow) present 4 in common, while that of PE (B) and RS (B) based rankings depict 4 communal among top ten predicted RSs. However, ranking varies for each RS bowler w.r.t. different metrics due to the incorporation of different statistical measures. For instance, M Morkel has got a maximum of 7^{th} rank by ICC in 2016 and he is on 1^{st} position, while Shakib Al Hasan acquired 3^{rd} rank by ICC and he is on 2^{nd} number as per ranked by RS (B) score. The reason is that M Morkel has played less matches, therefore, he has much better positively correlated feature scores as compared to Shakib Al Hasan. These features show a bitsy declining trend as more Co-bowlers are involved or more matches are played by underlying player. Similar causes are observed with the case of JP Faulkner but based on our detailed analysis, we predict that in the near future, he will get much higher rank by ICC as compared to his current rank. Nonetheless, all presented bowlers are RSs for sure as per validated by ICC rankings. Thus, we conclude that our incorporated metrics are robust for RSP.

VIII. CONCLUSION

Measures are explicitly adopted for rising star prediction in batting and bowling domains. More precisely, three categories (Co-players, Team and Opposite teams) are incorporated, in which 9 and 11 features are defined for the prediction of batting and bowling rising stars, respectively. Two types of datasets are generated based on weighted average and performance evolution metrics. The defined features are tested while employing generative (BN and NB) and discriminative (SVM

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

TABLE IV
BATTING RSS RANKING: RANKING COMPARISON OF TOP 10 RS BATSMEN BASED ON DEFINED THREE METRICS

Ranks	WA (B)			PE (B)			RS (B)		
	RS Name	Country	Highest ICC ranking 2013-16	RS Name	Country	Highest ICC ranking 2013-16	RS Name	Country	Highest ICC ranking 2013-16
1	V Kohli	India	1^{st}(2014)	V Kohli	India	1^{st}(2014)	AJ Finch	Australia	7^{th}(2015)
2	Hashim Amla	South Africa	1^{st}(2013)	Hashim Amla	South Africa	1^{st}(2013)	V Kohli	India	1^{st}(2014)
3	GJ Maxwell	Australia	7^{th}(2013)	Q de Kock	South Africa	3^{rd}(2016)	AM Rahane	India	23^{rd}(2016)
4	S Dhawan	India	5^{th}(2014)	JE Root	England	7^{th}(2016)	S Dhawan	India	5^{th}(2014)
5	MJ Guptill	New Zealand	5^{th}(2016)	MJ Guptill	New Zealand	5^{th}(2016)	Hashim Amla	South Africa	1^{st}(2013)
6	Q de Kock	South Africa	3^{rd}(2016)	S Dhawan	India	5^{th}(2014)	RG Sharma	India	5^{th}(2016)
7	PR Stirling	England	9^{th}(2013)	Umar Akmal	Pakistan	13^{th}(2013)	GJ Maxwell	Australia	7^{th}(2013)
8	RG Sharma	India	5^{th}(2016)	AJ Finch	Australia	7^{th}(2015)	Q de Kock	South Africa	3^{rd}(2016)
9	Umar Akmal	Pakistan	13^{th}(2013)	RG Sharma	India	5^{th}(2016)	F du Plessis	South Africa	9^{th}(2016)
10	JC Buttler	England	11^{th}(2016)	GJ Maxwell	Australia	7^{th}(2013) 3	DA Miller	South Africa	19^{th}(2015)

TABLE V
BOWLING RSS RANKING: RANKING COMPARISON OF TOP 10 RS BOWLERS BASED ON DEFINED THREE METRICS

Ranks	WA (Bow)			PE (Bow)			RS (Bow)		
	RS Name	Country	Highest ICC ranking 2013-16	RS Name	Country	Highest ICC ranking 2013-16	RS Name	Country	Highest ICC ranking 2013-16
1	M Morkel	South Africa	6^{th}(2016)	M Morkel	South Africa	6^{th}(2016)	MA Starc	Australia	1^{st}(2015)
2	Shakib Al Hasan	Bangladesh	3^{rd}(2015)	Shakib Al Hasan	Bangladesh	3^{rd}(2015)	M Morkel	South Africa	6^{th}(2016)
3	DW Steyn	South Africa	2^{nd}(2013)	R Ashwin	India	7^{th}(2013)	JP Faulkner	Australia	15^{th}(2015)
4	Junaid Khan	Pakistan	9^{th}(2014)	Junaid Khan	Pakistan	9^{th}(2014)	B Kumar	India	6^{th}(2014)
5	MA Starc	Australia	1^{st}(2015)	DW Steyn	South Africa	2^{nd}(2013)	DW Steyn	South Africa	2^{nd}(2013)
6	MJ McClenaghan	New Zealand	19^{th}(2015)	TG Southee	New Zealand	9^{th}(2015)	ST Finn	England	2^{nd}(2013)
7	SP Narine	West Indies	1^{st}(2013)	SP Narine	West Indies	1^{st}(2013)	R McLaren	South Africa	5^{th}(2014)
8	JC Tredwell	England	8^{th}(2014)	ST Finn	England	2^{nd}(2013)	M. Shami	India	10^{th}(2014)
9	ST Finn	England	2^{nd}(2013)	MA Starc	Australia	1^{st}(2015)	GH Dockrell	Ireland	21^{st}(2013)
10	TG Southee	New Zealand	9^{th}(2015)	MJ McClenaghan	New Zealand	19^{th}(2015)	M. Nabi	Afghanistan	20^{th}(2016)

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2682162, IEEE Access

and CART) machine learning algorithms. For batting domain, Co-batsmen category suppresses the remaining two categories, while in bowling realm, Team category outperforms for rising star prediction. Overall, it is observed that NB outperforms the remaining models. Finally, ranking lists of rising stars based on weighted average, performance evolution and rising star score are presented for both domain. These rankings are compared with the ICC rankings during 2013-16 and it is found that our presented approaches are functional for rising star prediction. Therefore, these features can also be used for rising star prediction in test and T20 formats. Moreover, some additional features such as opposite team diversity, home or away, 100s, 50s (for batsmen) and 4, 5 wickets (for bowlers) can also be incorporated in order to get even better results. Finding RSs within the cricket and other domains is quite useful, so that the authorities (coaches, managers etc.) can put efforts to maximize the expertise of such RSs in order to get the optimal performances in future. Similar methodology can be adopted for RSP in different sports domains and other organizations.

IX. Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (Nos. 61370194, 61411146001, 61502048).

References

[1] X.-L. Li, C. S. Foo, K. L. Tew, and S.-K. Ng, "Searching for rising stars in bibliography networks," in *International Conference on Database Systems for Advanced Applications*. Springer, 2009, pp. 288–292.

[2] A. Daud, M. Ahmad, M. Malik, and D. Che, "Using machine learning techniques for rising star prediction in co-author network," *Scientometrics*, vol. 102, no. 2, pp. 1687–1711, 2015.

[3] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1247–1252.

[4] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Temporal expert finding through generalized time topic modeling," *Knowledge-Based Systems*, vol. 23, no. 6, pp. 615–625, 2010.

[5] H. O. Stekler, D. Sendor, and R. Verlander, "Issues in sports forecasting," *International Journal of Forecasting*, vol. 26, no. 3, pp. 606–621, 2010.

[6] D. Beaudoin and T. B. Swartz, "The best batsmen and bowlers in one-day cricket," *South African Statistical Journal*, vol. 37, no. 2, p. 203, 2003.

[7] G. R. Amin and S. K. Sharma, "Measuring batting parameters in cricket: A two-stage regression-owa method," *Measurement*, vol. 53, pp. 56–61, 2014.

[8] A. Daud, F. Muhammad, H. Dawood, and H. Dawood, "Ranking cricket teams," *Information Processing & Management*, vol. 51, no. 2, pp. 62–73, 2015.

[9] G. Tsatsaronis, I. Varlamis, S. Torge, M. Reimann, K. Nørvåg, M. Schroeder, and M. Zschunke, "How to become a group leader? or modeling author types based on graph mining," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2011, pp. 15–26.

[10] A. Daud, R. Abbasi, and F. Muhammad, "Finding rising stars in social networks," in *International Conference on Database Systems for Advanced Applications*. Springer, 2013, pp. 13–24.

[11] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to platt's smo algorithm for svm classifier design," *Neural computation*, vol. 13, no. 3, pp. 637–649, 2001.

[12] W.-Y. Loh, "Classification and regression tree methods," *Encyclopedia of statistics in quality and reliability*, 2008.

[13] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[14] K. M. Leung, "Naive bayesian classifier," *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

[15] P. J. Bracewell, K. Ruggiero *et al.*, "A parametric control chart for monitoring individual batting performances in cricket," *Journal of Quantitative Analysis in Sports*, vol. 5, no. 3, pp. 1–19, 2009.

[16] S. Mukherjee, "Identifying the greatest team and captaina complex network approach to cricket matches," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 23, pp. 6066–6076, 2012.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[18] S. Mukherjee, "Quantifying individual performance in cricketa network analysis of batsmen and bowlers," *Physica A: Statistical Mechanics and its Applications*, vol. 393, pp. 624–637, 2014.

[19] S. Akhtar and P. Scarf, "Forecasting test cricket match outcomes in play," *International Journal of Forecasting*, vol. 28, no. 3, pp. 632–643, 2012.

[20] T. P. Sahu, N. K. Nagwani, and S. Verma, "Selecting best answer: An empirical analysis on community question answering sites," *IEEE Access*, vol. 4, pp. 4797–4808, 2016.

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.