Rising Cricketer Prediction Using Classification Models

Waqas Ahmad

665-FBAS/MSCS/S12

`Submitted in partial fulfilment of the requirements for the
MS Degree in Computer Science with
Specialization in Data Mining
At the Faculty of Basic and Applied Science,

INTERNATIONAL ISLAMIC UNIVERSITY ISLAMABAD

Supervisor:

Dr. Ali Daud

August, 2015

رَبِّ زِدْنِي عِلْمًا ۝

(طٰہٰ: 114)

اے میرے رب! میرے علم میں اضافہ فرما۔

In the name of

Allah,

The most Merciful and Compassionate the most Gracious and the Beneficent whose help and Guidance we always solicit at every step, and every moment.

Dedication

I dedicate this project to my beloved father ***Shams Ur Rehman.***

DECLARATION

I **WAQAS AHMAD** S/O **SHAMS UR REHMAN**

Registration No. **665-FBAS/MSCS/S12**

Student of MS in Computer Science at International Islamic University do hereby solemnly declared that the thesis entitled "**Rising Cricketer Prediction Using Classification Models**", submitted by me in partial fulfilment of MS degree in Computer Science, is my original work, except where otherwise acknowledged in the text, and has been submitted or published earlier and shall not, in future, be submitted by me for obtaining any degree from this or any other university or institution.

Student's Signature

Date: 04 August, 2015

DECLARATION

I Prof/Dr. **ALI DAUD**

Supervisor of student Mr. **WAQAS AHMAD**

do hereby solemnly declare that the thesis entitled "**Rising Cricketer Prediction Using Classification Models**" being submitted as partial fulfillment of M.S degree in the discipline of **Computer Science** has been completed under my guidance and supervision and is an original work of the student except where otherwise acknowledged in the text. It has not been submitted or published earlier for obtaining any degree from this or any other university or institution.

This thesis is completed in all respects and I am fully satisfied with the quality of student's research work. Now it is ready to be evaluated by external subject experts.

Date: 04 August, 2015 **Signature:** A Daud

Name in full: Dr. ALI DAUD

Address: Head of Data Mining and Information Retrieval Group, Assistant Professor, Islamic International University,

Islamabad

Phone: 051-9019509

Email: ali.daud@iiu.edu.pk

Department of Computer Science and Software Engineering

International Islamic University Islamabad

Date: 04/04/2015

Final Approval

This is to certify that we have read and evaluate the thesis entitled **"Rising Cricketer Prediction Using Classification Models"** submitted **by Waqas Ahmad** under **Reg No. 665-FBAS/MSCS/S12.** It is our judgment that this thesis is of sufficient standard to warrant its acceptance by International Islamic University, Islamabad for the degree of **MS in Computer Science**

COMMITTEE

External Examiner
Prof. Dr. Muhammad Afzal
Director,
Dr.A.Q Khan Institute of Computer Science and Information Technology,
Kahuta, Distt Rawalpindi.

Muhammad Afzal
4/4/15

Internal Examiner
Dr. Ayyaz Hussain
Assistant Professor DCS & SE,
International Islamic University Islamabad.

Supervisor
Dr. Ali Daud
Assistant Professor DCS & SE,
International Islamic Unviersity Islamabad.

A Daud

Acknowledgement

All praise to Almighty Allah, who gave me understanding, courage and patience to complete this research. I thank him to be always there for us and to give me the world best parents who encouraged me at every stage of life and due to their efforts I am at this position today. I also thank him to give me a very cooperative teacher and supervisor in this research **Dr. Ali Daud** who help and encourage me to complete MS research thesis.

Waqas Ahmad

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1 Finding Rising Stars.. 1

1.2 Difference between Expert Cricketer and Rising Cricketer............................... 1

1.3 Classification Models.. 1

1.3.1 Generative Models...2

1.3.2 Discriminative Models..3

1.4 Cricket..3

1.4.1 Game Formats..3

1.4.2 Teams..3

1.5 Research Contribution...4

2. LITERATURE REVIEW 5

2.1 Sports..5

2.2 Others sports..6

2.3 Classification Models For Prediction...7

2.3.1 Discriminative Models...7

2.3.2 Generative Models..7

2.4 Problem Statement 9

2.5 Research Objectives 10

3. RESEARCH METHODOLOGY 11

3.1 Pre-processing 11

3.2 Constructing Co-Players 11

3.3 Features Definition... 12

3.3.1 Co Batsmen Runns... 13

3.3.2 Co Batsmen Average 14

3.3.3 Co Batsmen Strike Rate 14

3.3.4 Co Bowlers Average 15

3.3.5 Co Bowlers Strike Rate 16

3.3.6 Co Bowlers Economy 17

3.3.7 Team Win/Loss Ratio 17

3.3.8 Team Average 19

3.3.9 Opposite Team Win/loss Ratio 21

3.3.10 Opposite Team Average 24

3.4 Weighted Average (Batsmen) 26

3.5 Weighted Average (Bowler) 26

3.6 Average Relative Increase 27

4. EXPERIMENT 28

4.1 Dataset 28

4.2 Performance Evaluation 28

4.2.1 Accuracy 28

4.2.2 Precision 29

4.2.3 Recall 29

4.2.4 F-Measures 29

4.3 Implementation Of Classification Models In WEKA 29

4.4 Results and Discussion 30

4.4.1 Individual Features Analysis 30

4.4.2 Model Wise Combined Features Analysis 36

4.4.3 Category Wise Analysis 42

4.5 Rising Stars 50

4.5.1 Rising Star Score (Batsmen) 50

4.5.2 Average Relative Increase (Batsmen) 52

4.5.3 Rising Star Score (Bowler) 53

4.5.4 Average Relative Increase (Bowlers) 54

5. CONCLUSION 56

REFERENCES 57

LIST OF FIGURES

Figure 1: Selecting Co-player for Player x 11

Figure 2: Features Precision Analysis Using 1st Dataset (Batsmen) 31

Figure 3: Features Recall Analysis Using 1st Dataset (Batsmen)...............................31

Figure 4: Features F-Measure Analysis Using 1st Dataset (Batsmen).......................32

Figure 5: Features Precision Analysis Using 2nd Dataset (Batsmen).......................33

Figure 6: Features Recall Analysis Using 2nd Dataset (Batsmen) 33

Figure 7: Features F-Measure Analysis Using 2nd Dataset (Batsmen)....................34

Figure 8: Features F-Measure Analysis Using 1st Dataset (Bowlers).....................35

Figure 9: Features F-Measure Analysis Using 2nd Dataset (Bowlers)....................36

Figure 10: Model Wise Precision Analysis Using 1st Dataset (Batsmen)................37

Figure 11: Model Wise Recall Analysis Using 1st Dataset (Batsmen)....................38

Figure 12: Model Wise F-Measure Analysis Using 1st Dataset (Batsmen)..............38

Figure 13: Model Wise Precision Analysis Using 2nd Dataset (Batsmen)...............39

Figure 14: Model Wise Recall Analysis Using 2nd Dataset (Batsmen)...................40

Figure 15: Model Wise F-Measure Analysis Using 2nd Dataset (Batsmen).............40

Figure 16: Model Wise F-Measure Analysis Using 1st Dataset (Bowlers)..............41

Figure 17: Model Wise F-Measure Analysis Using 2nd Dataset (Bowlers).............42

Figure 18: Category Wise Precision Analysis Using 1st Dataset (Batsmen)............43

Figure 19: Category Wise Recall Analysis Using 1st Dataset (Batsmen)................44

Figure 20: Category Wise F-Measure Analysis Using 1st Dataset (Batsmen)...........45

Figure 21: Category Wise Precision Analysis Using 2nd Dataset (Batsmen)............46

Figure 22: Category Wise Recall Analysis Using 2nd Dataset (Batsmen)................47

Figure 23: Category Wise F-Measure Analysis Using 2nd Dataset (Batsmen)...........48

Figure 24: Category Wise F-Measure Analysis Using 1st Dataset (Bowlers)............49

Figure 25: Category Wise F-Measure Analysis Using 2nd Dataset (Bowlers)...........50

LIST OF TABLES

Table 1: Feature Categories ODI Batsmen..12

Table 2: Feature Categories ODI Bowlers...12

Table 3: Win loss ratio involving player (DJ Barvo)..19

Table 4: Against South Africa involving Player (AB de Villiers)............................21

Table 5: Win loss ratio against South Africa involving player (AB de Villiers)...........22

Table 6: Against Pakistan involving player (Umar Gul).......................................23

Table 7: Win loss ratio against Pakistan involving player (Umar Gul)......................23

Table 8: Teams average against Pakistan involving player (Umar Gul).....................25

Table 9: Top Ten Predicted Rising Star (Batsmen) From Rising Star Score................51

Table 10: Top Ten Predicted Rising Star (Batsmen) From ARI Method....................52

Table 11: Top Ten Predicted Rising Star (Bowlers) From Rising Star Score...............53

Table 12: Top Ten Predicted Rising Star (Bowlers) From ARI Score........................54

LIST OF ABBREVIATIONS

ARI	Average Relative Increase
Avg	Average
CART	Classification and Regression Tree
Econ	Economy
ICC	International Cricket Council
IPL	Indian Premier League
LR	Linear Regression
MEMM	Maximum Entropy Markov Model
NB	Naïve Bayesian
ODI	One Day International
OT	Opposite Team
SR	Strike Rate
SVM	Support Vector Machine
TD	Temporal Dimension
Wkt	Wicket

ABSTRACT

In cricket community rising cricketers are players who have currently a low profile and considered to be a star in future, while a player having good profile is considered to be a star in cricket. Classification models (Discriminative and Generative) for finding rising stars in research community have been used by few researchers. However, no research has been done to predict rising cricketers using classification models. Here we propose to predict the rising cricketers in cricket by using classification modeling techniques. Rising cricketers forecasting mainly depend upon the performance of already star cricketers, the team and opposite team in which they are currently playing.

In this work, first we collect player's data (from Jan, 2000 to Dec, 2009) then we measure three classes of features (Co-Player, Team and Opposite Team) to identify rising stars in cricket by using classification models, rising star score algorithm and AVR (Average Relative Increase). The results show that Discriminative models (CART, SVM and MEMM) have best result comparatively Generative Models (Bayes net, Naïve Bayes). And finally we generate and present top ten predicted rising star list using rising star score and AVR (Average Relative Increase). Year 2011 ranking of player's data is being used to compare with our predicted rising star list.

CHAPTER 1

INTRODCUTION

1. INTRODUCTION

1.1 Finding Rising Stars

Finding the rising stars is an important but difficult human resource exercise in all organizations. Rising stars are those who at present have comparatively lower profiles but may be in the long run, considered to be a star in cricket. Searching for rising star in organizations a small amount of research work done in this area like (Xiao-Li, Foo, Tew, & Ng, 2009), Daud, A., Abbasi, R., & Muhammad, F. (2013). This research was based on social network. It is also applicable to find rising stars in different sports like baseball, basketball, football and cricket. Here we propose to find the rising stars in cricket. There are different statistical techniques and models being used for prediction. A predictive model is simply a mathematical function that is able to learn the mapping between a set of input data variables, usually bundled into a record, and a response or target variable (Alex Guazzelli, 2012). Particularly we will use the discriminative models to predict the rising cricketer.

1.2 Difference between Expert Cricketer and Rising Cricketer

Expert cricketer is the one who is already a star in cricket, having too much experience and huge profile in number while rising cricketer is said to be new in cricket and may be the future star by some aspects, he may be known as rising star or a rising cricketer. Name like Sachin Tendulkar (India), Kumar Sangakkara (Sri Lanka) and Hashim Amla (South Africa) are well known stars so in our terminology they will be expert cricketers. Quinton de Kock (South Africa) and Anwer Ali (Pakistan) having currently low profile but they might be rising cricketers.

1.3 Classification Models

Classification is a type of data analysis that extracts techniques relating key data classes to predict categorical class labels (Han, Kamber, & Pei, 2011). For example, a model can be built to categorize students as either computer literate or not, or a prediction model to predict the students final exam grades on the bases of their midterm exam score and previous exams scores. Researcher proposed many classification and prediction methods such as machine learning, pattern recognition and statistics. The classification techniques are further divided into two main categories that are:

- Generative Models
- Discriminative Models

1.3.1 Generative Models

"Generative models define a joint probability distribution *p*(**X**, **Y**) where **X** and **Y** are random variables respectively ranging over observation sequences and their corresponding label sequences" (Tang, Hong, Zhang, & Liang, 2007). In order to compute the conditional probability *p*(*y*|*x*), Bayesian rule is employed:

$$y = argmax_y p(y|x) = argmax_y \frac{p(x,y)}{p(x)} \qquad (1)$$

Bayes Net and Naïve Bayes are most usually used generative models that are defined in section 2.1.2.2.

1.3.2 Discriminative Models

Discriminative models are important class of probabilistic models with solid statistical foundation (Fang, Si, & Mathur, 2010). The author's (Tang, Hong, Zhang, & Liang, Information Extraction: Methodologies and Applications, 2007) defined discriminative models as conditional distribution p(y|x) of observation and label series. That means once isolating the most probable label succession for a certain observation succession, directly use the conditional distribution by the discriminative models, without any problem to make dependence supposition on observations or detail every feasible observation successions to calculate the trivial probability p(x).

Discriminative models have been used in the recent past in many machine learning applications, partly because of their remarkable theoretical features. Various discriminative models have been functional to various retrieval problems in the information retrieval field (Nallapati, 2004).

However, large research has been conducted to design classification models for expert search and information retrieval but no attempt has been made for rising stars prediction.

Data classification is a two-step process,

Model construction: Describing a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction: training set. The model is represented as classification rules, decision trees, or mathematical formulae.

Model usage: For classifying future or unknown objects. Estimate accuracy of the model. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model (Jiawei Han and Micheline Kamber, 2006)

1.4 Cricket

Cricket is a bat-and-ball game played between two teams of 11 players each. Each team takes its turn to bat, attempting to score runs, while the other team fields. Each turn is known as an innings. The bowler delivers the ball to the batsman who attempts to hit the ball with his bat far enough for him to run to the other end of the pitch and score a run. Each batsman continues batting until he is out. The batting team continues batting until ten batsmen are out, at which point the teams switch roles and the fielding team comes in to bat (Pankush Kalgotra, Ramesh Sharda and Goutam Chakraborty 2013).

1.4.1 Game Formats

International cricket has a variety of the game for everyone, from the five-day tactical tension of the Test match to the bat-swinging fireworks of Twenty20. Basically there are three types of basic formats in international cricket. 1. Test match Cricket 2. ODI (One Day International) and 3. Twenty20. Test match is of five continuous day's match of two innings for each side. The ODI and T20 are called limited over matches having only one innings per side. In ODI there are 50 overs per innings and 20 over per inning in T20 match formats (http://www.dummies.com/how-to/content/cricket-for-dummies-cheat-sheet.html).

1.4.2 Teams

There are three categories of country team membership in ICC (International Cricket Council): Full Members, Associate Members, and Affiliate Members. In the highest category, there are 10 Full Members. Below the Full Members are the 37 Associate Members. In the lowest category, there are 60 Affiliate Members. Full member team can play all three type of formats and the teams are Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri-

Lanka, West Indians and Zimbabwe. Top Associate members play ODI and T20 matches but they cannot play Test cricket. And all other members are those who follow the ICC rules for playing cricket (http://en.wikipedia.org/wiki/List_of_International_Cricket_Council_members).

1.5 Research Contribution

We have gone through the literature review techniques used for classification modeling for finding information retrieval and expert search, We found that rising start using classification models is used to find the future rising research authors stars but not in cricket or even in any other sports. Our objective is to use classification models for finding rising cricketers. The main task in the Social Network Mining is to discover rising stars in research community by using classification models to mining the researcher's social network in term of their co-author relationships (Daud, Abbasi, & Muhammad, 2013).

CHAPTER 2

LITERATURE REVIEW

2. LITERATURE REVIEW

Very little research has been done on predicting rising stars. There is no work done in sports field using the co-relation performance for prediction. Our proposed research related to prediction of cricketer in sports using discriminative models, so the literature review given below about players prediction in cricket and brief of the classification models.

2.1 Sports

Traditionally the batsmen and bowlers are rated on their batting or bowling average respectively. However in a game like Cricket it is always important the way one scores the runs or claims a wicket. Scoring runs against a strong bowling line or delivering a brilliant performance against a team with strong batting line deserves more credit. We explore the application of Social Network Analysis (SNA) to rate the players in a team performance (Satyam Mukherjee, 2013).

The success of a team (or captain) is determined by the 'quality' of wins and not on the number of wins .Author has used the diffusion based PageRank algorithm on the networks in order to measure the importance of winning a match by which rank the teams and its captain respectively(Satyam Mukherjee, 2012).

The study classifies the performance of all-rounders who participated in IPL (Indian Premier League) based on their strike rate and economy rate, where strike rate is defined as number of balls bowled divide by the number of wickets taken by a bowler and economy rate is the number of runs conceded per six balls respectively (Paul J.van Staden, 2008). Several predictor variables that are supposed to influence the performance of all-rounders are considered. Step-wise Multinomial Logistic Regression (SMLR) is used to identify the significant predictors, which were used to predict the expected class of an incumbent all-rounders using naïve Bayesian classification model (Hemanta Saikia and Dibyojyoti Bhattacharjee, 2011).

All-rounders i.e. players with the ability to both bat and bowl play a key role in cricket, whatever is the format of the game. The study measures the performance of all-rounders in Indian Premier Leagues (IPL) based on their strike rate and economy rate. A Naïve Bayesian classification model is developed that can use the significant predictors to predict the class in which an incumbent all-rounder is expected to lie. The classifier is build based on the performance of all-rounders who participated in IPL-I and II, and the validity of the classifier is subsequently tested over the incumbent all-rounders of IPL-III. ((Hemanta Saikia and Dibyojyoti Bhattacharjee, 2010).

2.2 Other sports

Not only is the game phase important (i.e., corner, free-kick, open-play, counter attack etc.), the strategic features such as defender proximity, interaction of surrounding players, speed of play, coupled with the shot location play an impact on determining the likelihood of a team scoring a goal (Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr and Iain Matthews, 2015). They use their spatiotemporal strategic features, which can accurately measure the likelihood of each shot. They use this analysis to quantify the efficiency of each team and their strategy.

Discriminative learning approach to automatically train models to predict near term game events given current game conditions. Building upon and combining discriminative learning techniques (such as Conditional Random Fields) along with techniques for spatial regularization and non-negative matrix factorization (Yisong Yue, Patrick Lucey, Peter Carr, Alina Bialkowski, Iain Matthews, 2014). They show how to influence basic high-level domain knowledge of sports gameplay to train accurate predictive models while automatically inferring an interpretable feature representation.

A Bayesian Network model for forecasting association football matches in which the subjective variables represent the factors that are important for prediction (A.C. Constantinou, N.E. Fenton, M. Neil, 2012). The model, which we call 'pi-football', generates predictions for a particular match by considering generic factors for both the home and away team, namely: 1) strength, 2) form, 3) psychology and 4) fatigue.

2.3 Classification Models for Prediction

Classification models are divided in two main categories such as generative and discriminative those are described below.

2.3.1 Discriminative Models

It has been showing that characteristic based discriminative models can constantly and widely do better than existing state of the art retrieval models with the exact selection of characteristics (Metzler & Croft, Linear feature-based models for information retrieval, 2007). Classification models are also used for Automatic Speech Recognition (ASR) problems. ASR can be defined as the independent, computer-driven transcription of spoken language into readable text in real time (Stuckless, 1994). Machine language paradigms are used to solve ASR problems (Li Deng and Xiao Li, 2013). They give a brief over view of discriminative and generative learning, supervised, unsupervised, active and adaptive learning and also Bayesian learning. Discriminative and generative classifiers are also used in robust 3D brain MRI segmentations problems (Liu, Iglesias, & Zhuowen, 2013). Outcomes of generative and discriminative learning on relationship between object and categories (Hsu & Griffiths, 2010).

2.3.1.1 Maximum Entropy Markov Model (MEMM)

The early work of affecting discriminative models in Information Retrieval (IR) track back to 1980s in which the maximum entropy methods was studied to get around term independence suppositions in probabilistic generative models (Cooper, Exploiting the Maximum Entropy Principle To Increase Retrieval Effectiveness, 2007). MEMM thinks about observation series to be conditioned upon rather than created through the label succession.

2.3.1.2 Support Vector Machine (SVM)

A SVM model is used for fault diagnosis in dynamic processes (Bolivar, Hidrobo, & G, 2013). Alex Rudnick and Michael Gasser used MEMM to solve the problem of inexpensive approach for building a large vocabulary lexical system for hybrid rule-based machine translation (RBMT) system (Rudnick & Gasser, 2013).

Support Vector Machines (SVM) is a supervised learning model with connected learning algorithms that investigate data and identify patterns. SVM (Nallapati, 2004) creates a hyper plane or set of hyper planes in a high dimensional space, which can be utilized for classification, regression, or other tasks.

2.3.1.3 Classification and Regression Tree (CART)

CART is basically a non-parametric learning approach that results in either regression or classification tree depending variables (features) are either categorical or numeric (Page, Ward, & Worrall-Carter, 2013) (Speybroeck, 2012). Fundamentally, the method of CART contains three parts (Loh, 2011). 1) Construction of maximum tree. 2) Selection of right tree size. 3) Classify new data using already constructed tree.

In a simple form, our aim is to predict a response or class 'y' from inputs vector ($\mathbf{X}_{1...}$ $\mathbf{X}_m$). A binary tree is then constructed; a test is performed on each internal node to create a left or right sub branch of tree. This process is repeated until leaf node is constructed.

2.3.2 Generative Models

The problem of learning classifiers from distributional data solve by using supervised and generative models (Lin, Lee, Bui, & Honavar, 2013). Generative techniques such as Bayesian Network and Naïve Bayes are used for solving identification of language varieties problem (Marcos, 2013).

2.3.2.1 Bayesian Network (Bayes Net)

Bayesian network is explained by (Neapolitan, 2003) and also look for detail (Wikipedia, 2013). Bayes Net is a directed acyclic graph (DAG) with edges as conditional dependencies and nodes as random variables in Bayesian prospective (Mascaro, Nicholso, & Korb, 2014). Consider a Bayes networks comprise of n nodes ($\mathbf{X}_1$.... $\mathbf{X}_n$). The joint probability density function of network are calculated as

$$P(\mathbf{X}_1 = x_1, \ldots\ldots, \mathbf{X}_n = x_n) = P(x_1,\ldots, x_n) = \prod_i P(x_i | x_1, \ldots, x_{i-1})$$

2.3.2.2 Naive Bayes

Naïve Bayesian model is described in detail by (Mitchell, 2010). The Naive Bayes is a probabilistic classification method that applies naïve (independence) hypothesis with Bayes algorithm among every pair of features. It can handle with both continuous and categorical independent variables and assumes that features are not statistically dependent on each other (Metsis, Androutsopoulos, & Paliouras, 2006). Given a class label y and a feature vector **X** = $\{x_1, x_2, \ldots\ldots\ldots\ldots, x_m\}$, The Bayes theorem is described as

$$P(y \mid X_1, X_2, X_3, \ldots, X_m) \propto P(X_1, X_2, X_3, \ldots, X_m \mid y)\ P(y)$$

Classification models have established increasing attention in Information Retrieval (IR), another related area, and learning to rank for IR, sparked interest among researchers in the community (Liu T. Y., March 2009). Most of the learning to rank models is discriminative in nature and they have shown improvements over their generative counterparts in ad hoc retrieval. LETOR (Liu, Xu, Qin, Xiong, & Li, Benchmark Dataset for Research on Learning to Rank for Information Retrieval, 2007) are reachable for research on learning to rank. Classification models are important class of probabilistic models they have given very useful results on expert search as claimed by (Fang, Si, & Mathur, 2010).

Although valuable work has been done on classification models for ad hoc retrieval and other IR domains, no such research has been achieved to design classification models for finding rising stars.

2.4 Problem Statement

Many organizations are concerned with identifying "rising stars. Commonly the identification process for rising cricketer or player selection in cricket uses purely the past performance and record of players for predicting the future rising cricketer. The main idea behind the co-players features is that if junior players that have played fewer matches and profile collaborate with senior players that have more matches and profile then a better chance to junior player to become a rising star in near future. The Classification modeling techniques are used in expert search and gives good result as compared to other techniques (Fang, Si, & Mathur, 2010) but no such attempt have been made to use classification algorithm for rising star prediction in cricket community. In this work focus is on finding the rising cricketer considering the features which we identify for different categories instead of player previous performance.

2.5 Research Objectives

- Collection of data of players
- Calculate the distinct features
- For finding rising cricketers among the players whom data is collected, we apply classification modeling techniques such as CART, Support Vector Machine (SVM), Maximum Entropy Markov Model (MEMM), Bayes Net and Naive Bayes.
- Predicting top 10 rising players using Rising Star Score and ARI (Average Relative Increase) methods.

CHAPTER 3

RESEARCH METHODOLOGY

3. RESEARCH METHODOLOGY

Dataset have been taken from ESPNCRICINFO (http://www.espncricinfo.com/statsguru). The data of Jan 2001- Dec 2009 is used to predict rising cricketer. Basically in rising cricketer predication we use ODI data with two categories bowling and batting.

3.1 Pre-processing

Following are the pre-processing steps, which are applied on data to get particular data for further implementation

- We eliminate all those players who played less than 10 matches. Only those players will be considered in the dataset who played at least 10 match.
- For batsmen dataset only those players are selected, whose batting number or position is 1 to 7. The tail of the team is mostly bowlers so to avoid bowler data as batsmen we eliminate the batting position of 8 to 11.
- For Bowlers bowling at order of 1-6 are selected for data set and after this the bowlers were not considered because bowlers after this are a part time bowler which may affect the data so these bowlers were also eliminated from dataset

3.2 Constructing Co-Players

For a player his co-players are those whom played with this player from the same team. Following figure shows that how to find the co-player for player x in figure.

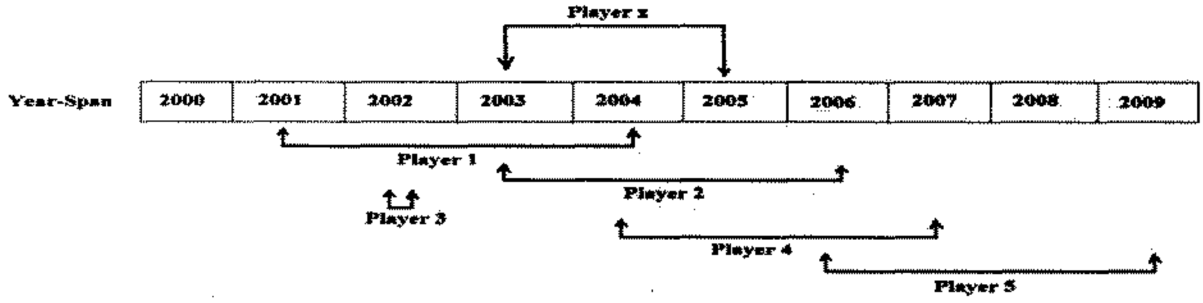


Figure 1: Selecting Co-player for Player x

Figure 1 show the year-span of ten years. Each player has own playing period in years. Figure1 we have to find the co-player from player 1 to 5 for player x as an example. Only those players are said to be the co-players of player x who played in any of year in which player x is played. So in figure player 1, player2 and player 4 are co-players of player of player x.

3.3 Features Definition

We take different features of each ODI which are batting and bowling categories. We calculate each feature one by one. And further each category feature is further divided into to three categories. Following two tables shows these categories and features.

Table 1: Feature Categories ODI Batsmen

C0-Batsmen	**Team**	**Opposite Team**
Co Batsmen Runs	Team Win/Loss Ratio	Opposite Team Win/Loss Ratio
Co Batsmen Average	Team Average	Opposite Team Average
Co Batsmen Strike Rate		

Table 1 shows the three categories that is co-batsmen, team and opposite team for batsmen dataset. Each category has different features in it like co-batsmen runs; co-batsmen average and co-batsmen strike rate are the features of co-batsmen category.

Table 2: Feature Categories ODI Bowlers

C0-Bowlers	**Team**	**Opposite Team**
Co Bowlers Average	Team Win/Loss Ratio	Opposite Team Win/Loss Ratio
Co Bowlers Strike Rate	Team Average	Opposite Team Average
Co Blowers Economy		

Similarly table 2 shows the bowling categories and its features. Co-bowler average, co-bowler strike rate and co-bowler economy are the features of co-bowler categories. Other two categories team and opposite team have same features as in ODI batsmen but dataset of bowler's team and opposite team are used.

3.3.1 Co Batsmen Runs

In all of the factors for calculating the ranking of batsmen in cricket runs is consider to be the first factor (http://www.icc-cricket.com/player-rankings). The sum of all co batsmen runs is said to be a player co batsman runs. After finding the co-player using the method shown in figure1 after that we add the runs of that co-players. Following is the expression for co-batsmen.

$$\text{Co Batsmen Runs (Player x)} = \sum_{i=1}^{n} \text{Co } P_i \qquad (2)$$

$$\text{Co Batsmen Runs (Player x)} = \text{Co } P_1 \text{ Runs} + \text{Co } P_2 \text{ Runs} + \text{Co } P_3 \text{ Runs} + \ldots \text{Co } P_n$$

For example

Player X Debut = 2003, Last =2005

And batsmen with runs of same team from year 2000 to 2009

1. Player 1 Runs scored =700
2. Player 2 Runs scored =856
3. Player 3 Runs scored =200
4. Player 4 Runs scored =900
5. Player 5 Runs scored =600

Player 1, Player 2 and player 4 are the only player whom played with player x at least once in their whole, so we take these batsman runs as co batsman runs so,

$$\text{Co Batsmen Runs (Player x)} = 700 + 856 + 900 \qquad (3)$$

$$\text{Co Batsmen Runs (Player x)} = 2456$$

3.3.2 Co Batsmen Average

Taking the average of all co-batsmen runs is said to be the co-batsmen average runs. The performance of player based on their strike rate and average (Paul J.van Staden, 2008). After getting co-player runs we find the average using following method

Co Batsmen Average (Player x) = (Co P_1 Avg+ Co P_2 Avg + Co P_3 Avg + …)/n (4)

Co P_1 Avg, Co P_2 Avg and Co P_3 Avg represents co-player 1, co-player 2, and co-player 3 respectively while n shows the number of co-players of player x.

For example

Co Batsmen Runs (Player x) = 700 + 856 + 900

Co Batsmen Average (Player x) = (700 + 856 + 900) / 3

Co Batsmen Average (Player x) = 818.67

Calculation shows that 818.67 is the co-batsmen average of player x.

3.3.3 Co Batsmen Strike Rate

Strike rate is defined as number of balls bowled divide by the number of wickets taken by a bowler. The performance of player based on their strike rate and average (Paul J.van Staden, 2008). Selecting the co-player for player x shown in figure 1 then we took those player strike rate and calculate them with following method

Co Batsmen S.R (Player x) = $\sum_{i=1}^{n}$ Co P_i S.R (5)

Co Batsmen S.R (Player x) = Co P_1 S.R+ Co P_2 S.R + Co P_3 S.R + … Co P_n S.R

Here S.R is strike rate while Co $P_{1,}$ Co $P_{2,}$ Co P_3 and Co P_4 represents co-player 1, co-player 2 and co-player 3 respectively.

For example

Player X Debut = 2003, Last =2005

And batsmen with strike rate of same team from year 2000 to 2009

1. Player 1 Strike Rate = 67.5
2. Player 2 Runs scored =99.0
3. Player 3 Runs scored =87.5
4. Player 4 Runs scored =103.5
5. Player 5 Runs scored = 78.0

Serial number 1, 2 and 4 players are said to be co-players of player x, so we take these batsman strike rate as co batsman strike rate

Co Batsmen S.R (Player x) = (67.5 + 99 + 103.5) / 3

Co Batsmen S.R (Player x) = 270 / 3

Co Batsmen S.R (Player x) = 90

Calculation shows that Co batsmen strike rate of player x is 90.

3.3.4 Co Bowlers Average

Procedure for selecting the co-player is same as discussed using figure 1. Average is one of the most important features that are used for raking the player. Here we calculate co bowler's average using following method where we consider co-bowlers playing with a player in same year.

$$\text{Co Bowlers Avg (Player x)} = \sum_{i=1}^{n} \text{Co } P_i \text{ Avg} \qquad (6)$$

$$\text{Co Bowlers Avg (Player x)} = \text{Co } P_1 \text{ Avg} + \text{Co } P_2 \text{ Avg} + \text{Co } P_3 \text{ Avg} + \ldots \text{Co } P_n \text{ Avg}$$

For example

Player x Debut = 2003, Last =2005 as in figure 1

And bowlers with debut and last detail with runs

1. Player 1 run concede = 3224 ,wickets taken =156
2. Player 2 runs concede = 2584,wickets taken =124
3. Player 3 runs concede = 865 ,wickets taken = 42
4. Player 4 runs concede = 2954,wickets taken =138
5. Player 5 runs concede = 2856, wicket taken =190

From example player 1, Player 2 and player 4 are the payers whom are the co-player of player x, so took these batsman runs as co batsman runs

Co Bowlers concede Runs (Player x) = 3224 + 2584 + 2954 = 8762

Co Bowlers Wickets (Player x) = 156 + 124 + 138 = 418

Co Bowlers Average (Player x) = 20.961

20.961 is the total co bowlers average for player x.

3.3.5 Co Bowlers Strike Rate

Strike rate plays important role in the ICC rating figures. After selecting co-players for blower we fetch their strike rates then we find the average of that co bowlers strike rates.

$$\text{Bowler S.R} = \frac{\sum \mathit{Balls\ Bowled}}{\sum \mathit{Wicket\ Taken}} \tag{7}$$

Equation defined as total numbers of ball bowled by a bowler per total number of wickets taken by a bowler which result in bowler strike rate mean that the how many balls he bowled to take as single wicket.

$$\text{Co Bowlers S.R (Player x)} = \sum_{i=1}^{n} \text{Co } P_i \text{ S.R} \tag{8}$$

$$\text{Co blowers S.R (Player x)} = \text{Co } P_1 \text{ S.R+ Co } P_2 \text{ S.R + Co } P_3 \text{ S.R + … Co } P_n \text{ S.R}$$

Equation was used to find the co-bowlers strike rate. Co P_1, Co P_2, and Co P_3 represents co-player 1, co-player 2 and co-player3 respectively.

For example

After selecting the co-bowler's se can find strike rate of each bowler and by taking the average of that value gives the co-bowlers strike rate.

Co-bowler from the above feature having data of ball bowled and wickets taken we have

Co Bowlers ball bowled (Player x) = 8647 + 4854 + 2954 = 16455

Co Bowlers Wickets taken (Player x) = 156 + 124 + 138 = 418

Co Bowlers Strike Rate (Player x) = 39.366

Co-Bowler's average shows 39.366 is the co-blower's strike rate

3.3.6 Co Bowlers Economy

Taking wickets is very important in cricket. Cricket rankings give more credit to a bowler for economy (http://www.relianceiccrankings.com). Following is the equation is for calculating bowler's economy

$$\text{Bowler's Economy} = \frac{\sum RG}{\sum OB} \tag{9}$$

In the equation $\sum$RG means total number of runs given while $\sum$OB represents total numbers of over bowled.

Co-Player method is used to find and select co-player as discussed in figure. Then we use the following expression to calculate co-bowlers economy.

$$\text{Co Bowlers Econ (Player x)} = \sum_{i=1}^{n} \text{Co } P_i \text{ Econ} \tag{10}$$

$$\text{Co blowers Econ (Player x)} = \text{Co } P_1 \text{ Econ} + \text{Co } P_2 \text{ Econ} + \text{Co } P_3 \text{ Econ} + \ldots \text{ Co } P_n \text{ Econ}$$

For example

Selecting the co-bowler's from the feature of co-bowler average with the additional data of number of runs given by co-bowlers and number of over bowled by a bowler.

Co Bowlers of player x Given Runs = 1224 + 2584 + 1954 = 5762

Co Blowers of player x Over bowled (P) = 350 + 419 + 315 = 1080

For economy we use the equation …

Bowlers economy = 5762/1080

So the co-bowler economy is

Co Batsmen Economy = 5.335

3.3.7 Team Win/Loss Ratio

The success of a team (or captain) is determined by the 'quality' of wins (Satyam Mukherjee, 2012). Team win/loss ratio is defined as win/loss ratio of team but here we calculate team win/loss for player and we say that those win/loss ratio of team in which matches particular player is played.

We calculate team win/loss ratio for two different situations

First method is use when Loss=0 then

$$\text{W/L Ratio} \quad = \quad Win * \frac{TM-(NR+Tie)}{Win+Loss} \tag{11}$$

And for all other cases method is

$$\text{W/L Ratio} \quad = \quad \frac{Win}{Loss} * \frac{TM-(NR+Tie)}{Win+Loss} \tag{12}$$

In Equations W/L and TM represents win/loss and total number of matches. NR represent the matches with No Result and Tie matches are those which result in draw at the end mean here the same with no team wins that match.

Example (Batsmen)

For calculating team win loss ratio for a particular player. Here we take GC Smith as an example to find the team win/loss for South Africa for only those matches in which this particular player GC Smith was played

$$\text{W/L Ratio (GC Smith)} \quad = \quad \frac{Win}{Loss} * \frac{TM-(NR+Tie)}{Win+Loss}$$

By putting the data in this equation

$$\text{W/L Ratio (GC Smith)} \quad = \quad \frac{81}{55} * \frac{143-(5+2)}{81+55} = 1.47$$

So in the upper example total number Matches = 143, Win = 81, Lost = 55, NR (No result) = 5 and Tie = 2 resulting the team win/loss ratio for player GC Smith is 1.47.

Example (Bowlers)

For calculating team win loss ratio of particular player. Here we take DJ Bravo from West Indies as an example to find the team win/loss for West Indies for only those matches in which this particular player DJ Bravo was played

$$\text{W/L Ratio (DJ Bravo)} = \frac{Win}{Loss} * \frac{Total\ Number\ of\ Matches-(NR+Tie)}{Win+Loss}$$

Following data variables are required for calculating W/L ratio (DJ Bravo)

Team: West Indies

Opposite Teams: Australia, Bangladesh, England, Sri Lanka, South Africa, Pakistan, India
New Zealand, Zimbabwe.

Table 3: Win loss ratio involving player (DJ Barvo)

Team	Mat	Won	Lost	Tied	NR
West Indies	98	42	50	0	6

By putting the data in this equation

W/L Ratio (DJ Bravo) = $\frac{42}{50} * \frac{98-(6+0)}{42+50} = 0.84$

3.3.8 Team Average

As a team win loss ratio has impact on a player similarly team average has effect on player performance. Team average runs ratio is define as average runs of team played its overall matches but here we calculate team average runs for player and we say that those average runs of team in which matches particular player is played.

We calculate simple team average runs as:

$$\text{Team Avg} = \frac{\sum TR}{\sum WF} \tag{13}$$

$\sum$TR and $\sum$WF represents total number of runs and total number of wicket fall respectively.

Example (Batsmen)

We calculate RG Sharma from team India as an example for his involving matches for India and team average in those particular matches.

$$\text{Team Average (RG Sharma)} = \frac{\sum TR}{\sum WF}$$

By putting the required data in the above equation

Team Average (RG Sharma) $= \frac{9070}{263} = 34.48$

Equation shows the 34.48 as Team average for team India only those matches in which RG Sharma as played while the 9070 is total team runs and 263 is the fall of wickets for team India.

Example (Bowlers)

Team average runs for a bowlers is quite different as compare to team average for batsmen following equation is use to find the team average for bowler.

$$\text{Team Avg} \quad = \quad \frac{\sum RGT}{\sum WTT} \qquad (14)$$

Expression shows that as total runs given by total number of wickets taken by team

We use the data of same player DJ Bravo team average as an example

Variable we use for calculating

Team: West Indies

Opposite Team: Australia, Bangladesh, England, Sri Lanka, South Africa, Pakistan, India

New Zealand, Zimbabwe.

Involving Player DJ Bravo

Total Runs Given by Team West Indies = 23520

Total Number of Wicket taken by team West Indies = 720

By putting the above values in the equation (3.14)

$$\text{Team Avg} \quad = \quad \frac{23520}{720} = 32.67$$

Result shows the team average of team (West Indies having DJ Bravo Played) per wicket means that conceding 32.67 runs team is able to take a single wicket.

3.3.9 Opposite Team Win/Loss Ratio

Opposite Team Win/Loss ratio is just like team win/loss ratio but here we calculate the win/loss ratio of team against which player is played. This feature is also that much important as team/loss ratio it has very high to its impact on a player performance.

We calculate team win/loss ratio for two different situations. First method is use when Loss=0

$$\text{OT W/L Ratio} = \sum OT\ Wins * \frac{\sum TM-(NR+Tie)}{\sum OTWins+\sum OT\ Loss}/\sum T \quad (15)$$

While in all other cases the following method was used

$$\text{OT W/L Ratio} = \frac{\sum OTWins}{\sum OT\ Loss} * \frac{\sum TM-(NR+Tie)}{\sum OTWins+\sum OT\ Loss}/\sum T \quad (16)$$

Equation calculates the Opposite team win/loss ratio in which OT represents Opposite Team while NR represents the matches with No Result.

For example (Batsmen)

Here we use opposite team win/loss ratio for team South Africa and player AB de Villiers as an example. First of all we find the following data of South Africa team, and only those matches which were played by AB de Villiers.

Table 4: Against South Africa involving Player (AB de Villiers)

Team	Mat	Won	Lost	Tied	Draw
Australia	12	8	3	0	1
England	11	4	3	0	4
India	6	2	3	0	1
Sri Lanka	2	2	0	0	0
Pakistan	5	1	3	0	1
West Indies	7	1	4	0	2
Bangladesh	4	2	4	0	0
New Zealand	5	1	4	0	1
Zimbabwe	2	1	2	0	0

After this we calculate using W/L Ratio Equation which is follow calculate the win/loss ratio of each team which is

Table 5: Win loss ratio against South Africa involving player (AB de Villiers)

Team	**W/L**
Australia	2.666
England	1.333
India	0.666
Sri Lanka	2
Pakistan	0.333
West Indies	0.25
Bangladesh	0.5
New Zealand	0.25
Zimbabwe	0.5

By taking the Sum of all team win/loss ratios and taking average we got the value

OT W/L Ratio = 0.944

Which is the opposite team win/loss ratio for team South Africa and particular those matches in which AB de Villiers was played.

Example (bowlers)

Here we use opposite team win/loss ratio for team Pakistan and player Umar Gul as an example. First of all we find the following data of Pakistan team and only those matches in which player Umar Gul was involved

Team: Pakistan

Player: Umar Gul

Table 6: Against Pakistan involving player (Umar Gul)

Team	Mat	Won	Lost	Tied	NR
Australia	5	3	2	0	0
Bangladesh	10	2	10	0	0
England	1	2	1	0	0
India	12	7	5	0	0
New Zealand	7	5	2	0	0
South Africa	7	4	3	0	0
Sri Lanka	10	4	6	0	0
West Indies	8	1	7	0	0
Zimbabwe	3	1	3	0	0

After this we calculate using W/L Ratio Equation which is follow calculate the win/loss ratio of each team which is

Table 7: Win loss ratio against Pakistan involving player (Umar Gul)

Team	W/L
Australia	1.5
Bangladesh	0.2
England	2
India	1.4
New Zealand	2.5
South Africa	1.333
Sri Lanka	0.666
West Indies	0.142
Zimbabwe	0.333

By taking the Sum of all team win/loss ratios and taking average we got the value

OT W/L Ratio = 1.119

Which is the opposite team win/loss ratio for team Pakistan and particular those matches in which Umar Gul was played.

3.3.10 Opposite Team Average

Opposite team average is an important factor that has impact on player's performance. Like for a bowler if the opposite team batting average is good then bowlers from his team may have weak bowling line. And similarly for batsmen have well opposite average runs means batting line of that team is weak. So that's why this feature was included.

Equation to find the opposite team average is different for batsmen and bowlers, so both the equations are defined in following example separately.

Example: (Batsmen)

Following equation was use to find the opposite team average

$$\text{OT Avg} = \frac{\sum OTR}{\sum WF} \tag{17}$$

Opposite Team average ratio is define as runs of opposite team average its overall matches but Here we calculate opposite team average for player and we say that those average runs per wicket of Opposite team in which matches particular player is played.

In the expression ∑OTR and ∑WF represents total number of runs and total number of wickets fall respectively.

Here for instants we take data of team South Africa and player AB de Villiers as an example. First of all we find the following data of South Africa team and only that matches in which player AB de Villiers as involved Data. Involve the number of average runs per wicket by each team against South Africa and then simply we take the average of all team so we got

OT Average (South Africa, AB de Villiers) = **31.901**

Example: (Bowlers)

Calculating the Opposite team average runs for a bowlers is totally different as compare to opposite team average for batsmen following equation is use to find the team average for blower.

Opposite Team average is define as runs of opposite team average score its overall matches but Here we calculate opposite team average for player and we say that those average runs per wicket of Opposite team in which matches particular player is played.

We calculate simple team average runs as:

$$\text{OT Average} = \frac{\sum OTCR}{\sum OTWT} \qquad (18)$$

In expression ∑OTCR is use for opposite team conceded runs while ∑OTWT is for opposite team wickets taken.

Here once again we use the same data set as we use in opposite team win/loss ratio. For team Pakistan and player Umar Gul as an example. First of all we find the following data of Pakistan team and only those matches in which player Umar Gul was involved Data. This is

Table 8: Teams average against Pakistan involving player (Umar Gul)

Team	Average
Australia	28.8
Bangladesh	43.45
England	26
India	38.83
New Zealand	24.36
South Africa	27.1
Sri Lanka	28.09
West Indies	33.45
Zimbabwe	42.05

Involve the number of average runs per wicket by each team against Pakistan and then simply we take the average of all team so we got

OT Average (Pakistan, Umar Gul) = **31.901**

3.4 Weighted Average (Batsmen)

For applying models on dataset one variable of dataset is required for threshold, so for this we take three main variables in batsmen data set. These are runs, average and strike rate. Give weightage of 33 to each factor

For example:

Player MJ Clarke Runs=4945, Avg=42.62 and SR= 77.64

Weighted Average = (33*runs + 33*Avg + 33* SR)/3

Weighted Average = (33*4945 + 33*42.62 + 33* 77.64)/3 = 55717

3.5 Weighted Average (Bowler)

Number of wickets, economy of bowler and strike rate of bowler are the three factor use weighted average. In bowling wicket have more importance as compare to economy and strike rate that's why we give 50 % to weightage to wickets and 25 % each to economy and strike rate.

For example:

Player Z khan Wickets= 219, econ= 4.92, and 36.8

Weighted Average = (50*wickets + 25*econ + 25* SR)/3

Weighted Average = (50*219 + 25*4.92 + 25* 36.8)/3 = 3997

3.6 Average Relative Increase

Two datasets we took same number of players who have highest Average Relative Increase in Runs (ARIR) for batsmen and Average Relative Increase in Wickets (ARIW) for bowlers. The notion for ARI is derived in the same way as (Tsatsronis, et al, 2011). The ARI is calculated as:

$$Relative\ Increase = \frac{current\ value - orignal\ value}{orignal\ value} \quad (19)$$

Example

Player with 5 years career
year 1 = 300, year 2 = 500, year 3 = 700, year 4 = 900, year 5 = 1100

$$Relative\ Increase = \frac{500-300}{300} = 0.67, \frac{700-500}{500} = 0.4, \frac{900-700}{700} = 0.29, \frac{1100-900}{900} = 0.23$$

If we multiply Relative Increase by 100 it's gives us percentage of increase e.g. 0.67*100 = 67% increase.

To calculate player average Relative increase is just to sum its Relative increase in 5 years and divided by the number of Relative increase in a specific time span.

$$ARI = \frac{67+40+29+23}{4} = \frac{159}{4} = 39.75\%$$

CHAPTER 4

EXPERIMENT

4. EXPERIMENT

4.1 Dataset

Dataset ranging from 2000 to 2009 have been taken from cricinfo to predict rising stars separately for both batsmen and bowlers. In dataset for batsmen player name, Debut, Last, Matches, Runs, Average and strike Rate were considered as data variables while Player name, Debut, Last, Matches, Wickets, Average, Economy and strike Rate were considered as data variables in dataset for bowlers. We made four types of dataset. In first two type for batting and bowling datasets we have taken top 150 runs and wicket takers from espncricinfo (www.cricinfo.org) that are called rising stars and also took 150 player having low profile that are called not rising stars. We used randomly 100 highly profile players and 100 lower profile players for testing and training.

4.2 Performance Evaluation

We used 5-fold cross validations method on 60 highly weighted average and 250 low weighted average data that were collected and calculated from ESPN cricinfo. The 5-fold cross validation means that the model (sample) set is divided into five equal parts. Four parts are used for training purpose while one part is used for testing. This process is repetitive five times and each time different five sample parts are used for testing. Then the average result rate is taken. After applying five-fold cross validation method on dataset, we check accuracy, precisions, and recall and f-measures on the results.

4.2.1 Accuracy

Accuracy of a dimension system is the extent of proximity of dimensions of a number to that numbers real value (Mitchell, GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION, 2010). In other words, the accuracy is how close the measured values to the actual (true) values. Accuracy is used to calculate of how binary classification experiment is recognized. Accuracy is a part of accurate results in a population for both true positive and true negative. It is the test parameters (Powers, 2007).

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn} \qquad (26)$$

tp stands for true positives tn stands for true negatives

fp stands for false positives fn stands for false negatives

4.2.2 Precision

The precision of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results. In other words, Precision is how close the measured values are to each other.

Precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives) (Powers, 2007).

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \qquad (28)$$

4.2.3 Recall

Recall is a measure of the ability of a prediction model to select instances of a certain class from a dataset. It is usually called sensitivity, and corresponds to the true positive rate. It is defined by the formula (Powers, 2007):

$$Recall = sensitivity = \frac{true\ positive}{(true\ positive + false\ negative)} \qquad (29)$$

True positive + false negative is total no. of analysis instances of the measured class.

4.2.4 F-Measures

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score (Powers, 2007):

$$\mathrm{F} = 2.\frac{\mathrm{precision.recall}}{\mathrm{precision + recall}} \qquad (30)$$

This is also known as the F1 measure, because recall and precision are evenly weighted.

4.3 Implementation of Classification Models in WEKA

The Waikato Environment for Knowledge Analysis (WEKA) is renowned as a most commonly used tool for research purpose in Data Mining and has reached extensive acceptance in the academic circles and industries (Witten & Eibe Frank, 2011). Weka is a set of machine learning and data mining algorithms. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka controls tools for classification, regression, data pre-processing, clustering, association rules, and visualization. Weka is open source and freely available software issued under the GNU (General Public License)

4.4 Results and Discussion

This section provides the detailed results of classification modeling techniques that we have been used for finding rising cricketers. We have also performed category wise and features wise results discussion. Results are basically in two main categories that are batting and bowling.

4.4.1 Individual Features Analysis

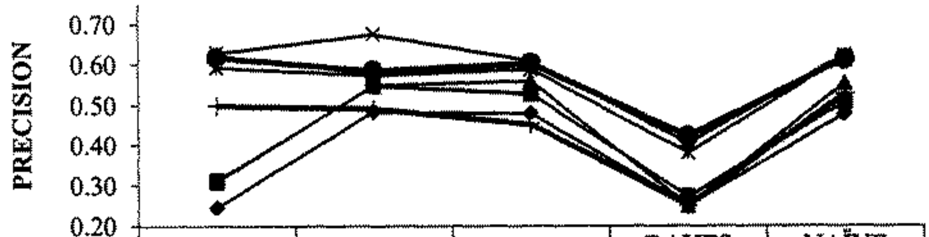
We have calculated precision, recall and f-measures of all features that we have described in chapter 3 by using classification models such as CART, MEMM, SVM, Bayes Net and Naïve Bayes on two types of dataset of both bowlers and batting. In the first type of dataset, player are selected for performance analysis on the basis of weighted average and in the second type of dataset, players are selected for performance analysis on the basis of Average Relative Increase (ARI) that described above in section 4.1.

Scenario 1: Individual Features Analysis Using Classification Models Based on Weighted Average Dataset (Batsmen)

In this section we analyze precision, recall and f-measure of batsmen individual features by using 1st dataset which is based on Weighted Average (runs, average and strike rate) described in the last section of chapter 3 on data sample of 10 to 100 players (batsmen).

Result: Precision Analysis of Features

Figure 2 show the precision analysis result of feature by using classification modeling techniques. In figure 2, we noted that Team Average Runs gives highest average result of accuracy 63%, 68%, 61%, 41% and 62% using CART, SVM, MEMM, BAYES NET and NAÏVE BAYES algorithms. Then Opposite Team Average Runs gives second highest average result of accuracy 62%, 59%, 60%, 42% and 61% using five classification algorithms. Discriminative nature classifiers give better accuracy result as compare to generative classifiers.

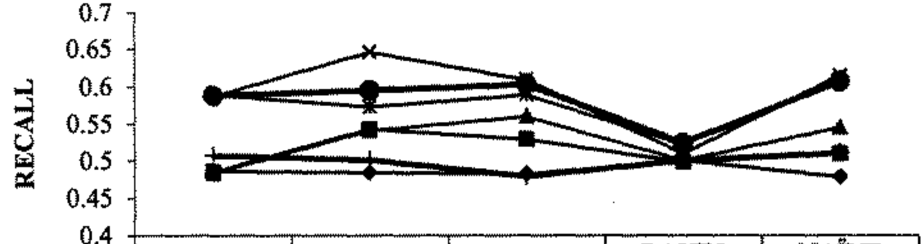


	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
—◆— Co-Bat Runs	0.25	0.48	0.48	0.25	0.48
—■— Co-Bat Avg	0.31	0.55	0.53	0.27	0.51
—▲— Co-Bat SR	0.31	0.55	0.56	0.25	0.55
—×— Team Avg Runs	0.63	0.68	0.61	0.41	0.62
—✱— Team W/L Ratio	0.59	0.57	0.59	0.38	0.62
—●— Opp T Avg Runs	0.62	0.59	0.60	0.42	0.61
—+— Opp T W/L Ratio	0.50	0.49	0.45	0.25	0.53

Figure 2 : Features Precision Analysis Using 1st Dataset (Batsmen)

Result: Recall Analysis of Features

Similarly figure 3 show the recall result of features using classification models on the data sample that were consisted of 10 to 100 players. In figure 3, we note that Team Average Runs gives highest accuracy of 64% using SVM algorithm. Its shows that in both precision and recall SVM Model give better result as compare to other models.

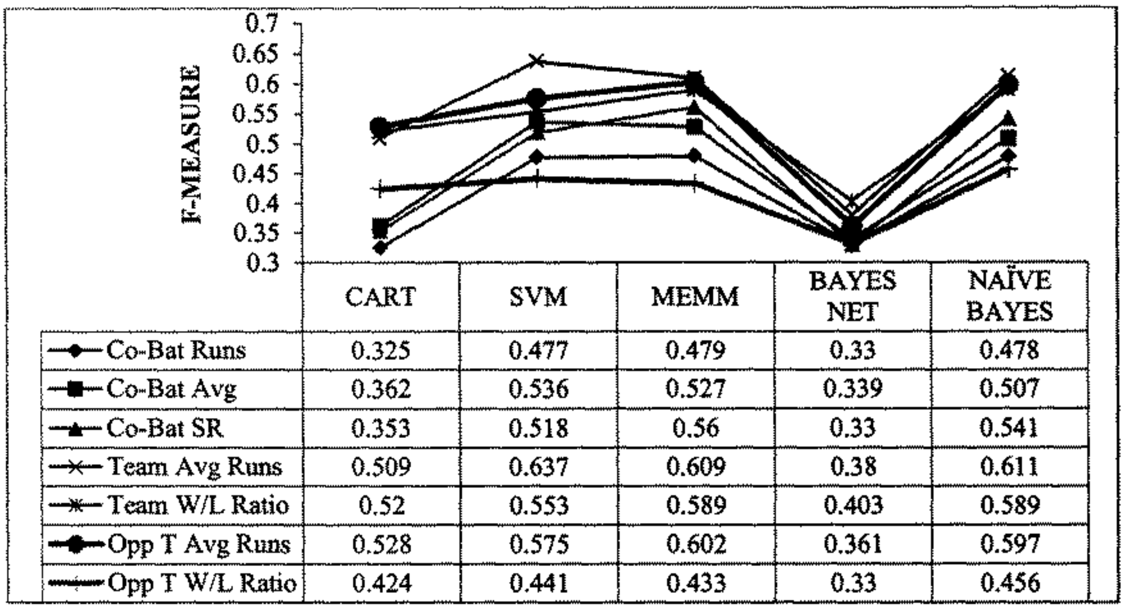


	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
—◆— Co-Bat Runs	0.486	0.484	0.482	0.5	0.478
—■— Co-Bat Avg	0.485	0.543	0.529	0.498	0.509
—▲— Co-Bat SR	0.483	0.541	0.56	0.5	0.545
—×— Team Avg Runs	0.587	0.646	0.609	0.509	0.614
—✱— Team W/L Ratio	0.589	0.573	0.589	0.522	0.605
—●— Opp T Avg Runs	0.588	0.595	0.603	0.524	0.606
—+— Opp T W/L Ratio	0.507	0.501	0.479	0.5	0.51

Figure 3: Features Recall Analysis Using 1st Dataset (Batsmen)

Result: F-Measure Analysis of Features

Figure 4 show the F-Measure result of features by using classification models on weighted average dataset of batsmen. Graph in figure 4 shows that Team Average Runs shows highest position with the highest accuracy of 63% which is using SVM model. And in other models it gives same result comparing with Opposite Team average Runs.



	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Bat Runs	0.325	0.477	0.479	0.33	0.478
■ Co-Bat Avg	0.362	0.536	0.527	0.339	0.507
▲ Co-Bat SR	0.353	0.518	0.56	0.33	0.541
× Team Avg Runs	0.509	0.637	0.609	0.38	0.611
✳ Team W/L Ratio	0.52	0.553	0.589	0.403	0.589
● Opp T Avg Runs	0.528	0.575	0.602	0.361	0.597
+ Opp T W/L Ratio	0.424	0.441	0.433	0.33	0.456

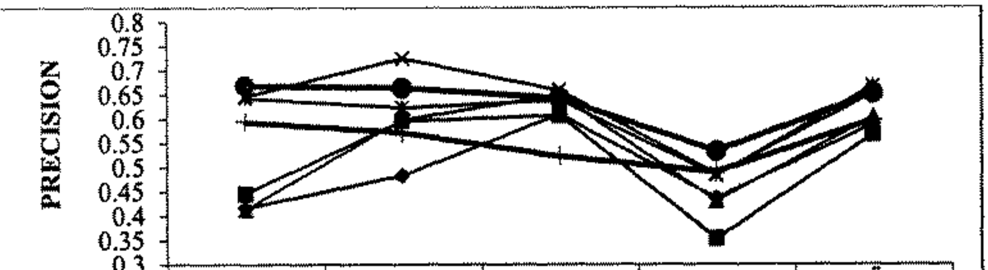
Figure 4: Features F-Measure Analysis Using 1st Dataset (Batsmen)

Scenario 2: Individual Features Analysis Using Classification Models Based on ARI (Average Relative Increase) Dataset (Batsmen)

In this section we analyze precision, recall and f-measure of batsmen features by using 2nd dataset Average Relative Increase (ARI) on data sample of 10 to 100 players (batsmen).

Result: Precision Analysis of Features

Figure 5 show the precision result of features by using 2nd dataset (ARI) Average Relative Increase that consist of 10 to 100 players (batsmen). In figure 5, we noted that Team Average Runs gives highest average result of accuracy 64%, 72%, 65%, 48% and 65% using CART, SVM, MEMM, BAYES NET and NAÏVE BAYES models respectively. Then Opposite Team Average Runs gives second highest average result of accuracy 66%, 66%, 64%, 53% and 65%.

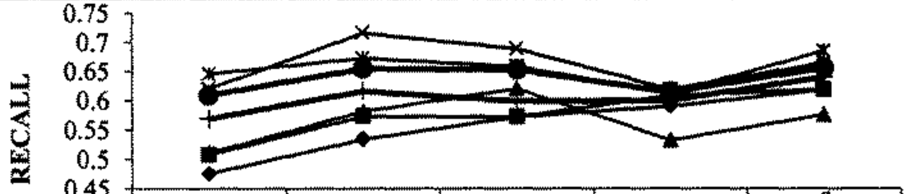


	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Bat Runs	0.416	0.483	0.61	0.435	0.588
■ Co-Bat Avg	0.445	0.596	0.608	0.352	0.567
▲ Co-Bat SR	0.413	0.599	0.65	0.4295	0.604
× Team Avg Runs	0.647	0.725	0.659	0.485	0.652
✱ Team W/L Ratio	0.643	0.624	0.639	0.482	0.667
● Opp T Avg Runs	0.668	0.664	0.643	0.533	0.652
+ Opp T W/L Ratio	0.594	0.572	0.523	0.49	0.597

Figure 5: Features Precision Analysis Using 2nd Dataset (Batsmen)

Result: Recall Analysis of Features

Similarly figure 6 show the recall result of features using classification models provide mostly the same result it gives in its precision in figure 5. Figure 6 shows that Team Average Runs gives better result using SVM models of 71 % and 68% with MEMM model. Team Win Loss Ratio gives second highest result on model CART and SVM algorithm that is 64% and 67% respectively.

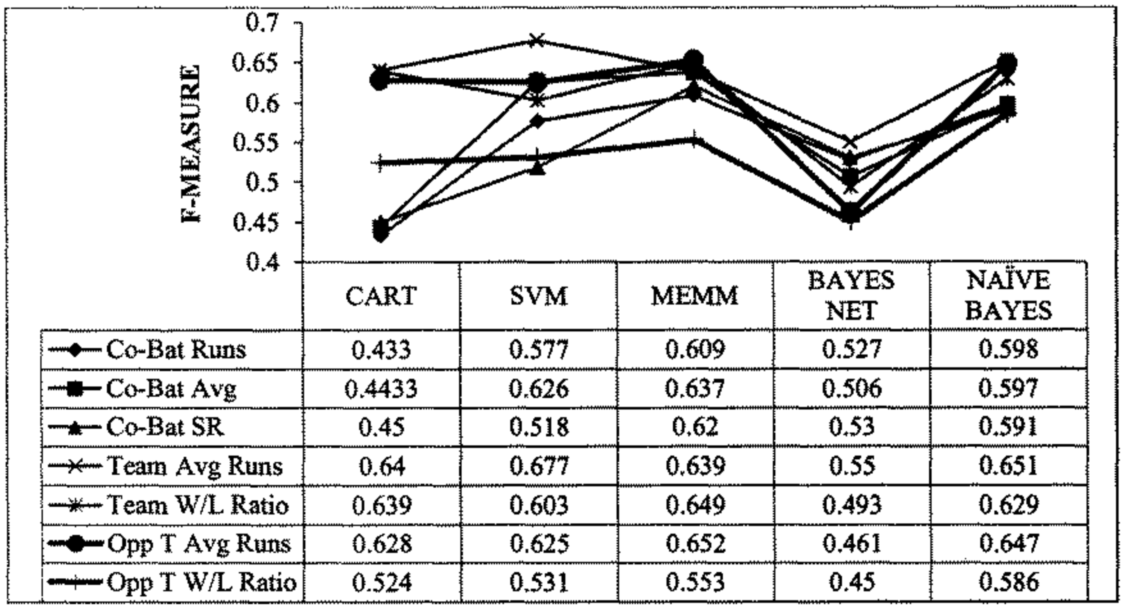


	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Bat Runs	0.476	0.534	0.572	0.59	0.618
■ Co-Bat Avg	0.508	0.573	0.572	0.608	0.619
▲ Co-Bat SR	0.512	0.581	0.62	0.532	0.575
× Team Avg Runs	0.62	0.716	0.689	0.619	0.664
✱ Team W/L Ratio	0.646	0.673	0.659	0.612	0.685
● Opp T Avg Runs	0.609	0.655	0.653	0.614	0.656
+ Opp T W/L Ratio	0.569	0.616	0.599	0.6	0.64

Figure 6: Features Recall Analysis Using 2nd Dataset (Batsmen)

Result: F-Measure Analysis of Features

Figure 7 show the F-Measure result of features by using classification models on Average Relative Increase (ARI) dataset of batsmen. Graph in figure 7 shows that Team Average Runs shows highest position with the highest accuracy of 67% which is using SVM model. And Team Average Runs feature in other models it gives same result comparing with Opposite Team average Runs feature.



	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Bat Runs	0.433	0.577	0.609	0.527	0.598
■ Co-Bat Avg	0.4433	0.626	0.637	0.506	0.597
▲ Co-Bat SR	0.45	0.518	0.62	0.53	0.591
× Team Avg Runs	0.64	0.677	0.639	0.55	0.651
* Team W/L Ratio	0.639	0.603	0.649	0.493	0.629
● Opp T Avg Runs	0.628	0.625	0.652	0.461	0.647
+ Opp T W/L Ratio	0.524	0.531	0.553	0.45	0.586

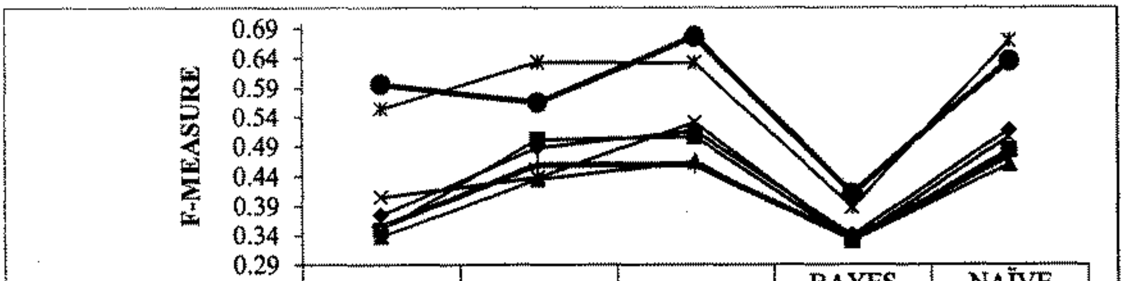
Figure 7: Features F-Measure Analysis Using 2nd Dataset (Batsmen)

Scenario 3: Individual Features Analysis Using Classification Models Based on Weighted Average Dataset (Bowlers)

In this section we analyze f-measure of bowlers feature that are described in chapter 3 by using 1st dataset of Weighted Average (Runs, Average and Strike rate) which is describe in last section of chapter 3.

Result: F-Measure Analysis of Features

Figure 8 shows the recall result of features using classification models provide highest accuracy result of feature Opposite Team Average Runs. In Figure 8, Team Average Runs gives better of 67% using MEMM model. Team Win Loss Ratio gives second highest result and same using SVM (Support Vector Machine) and MEMM (Maximum Entropy Markov Model) algorithms that are 63%.



	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Bat Runs	0.375	0.488	0.516	0.34	0.517
■ Co-Bat Avg	0.349	0.501	0.506	0.33	0.483
▲ Co-Bat SR	0.338	0.435	0.464	0.33	0.459
× Team Avg Runs	0.405	0.438	0.53	0.331	0.505
✳ Team W/L Ratio	0.554	0.634	0.632	0.387	0.669
● Opp T Avg Runs	0.596	0.565	0.676	0.41	0.635
+ Opp T W/L Ratio	0.355	0.459	0.459	0.334	0.473

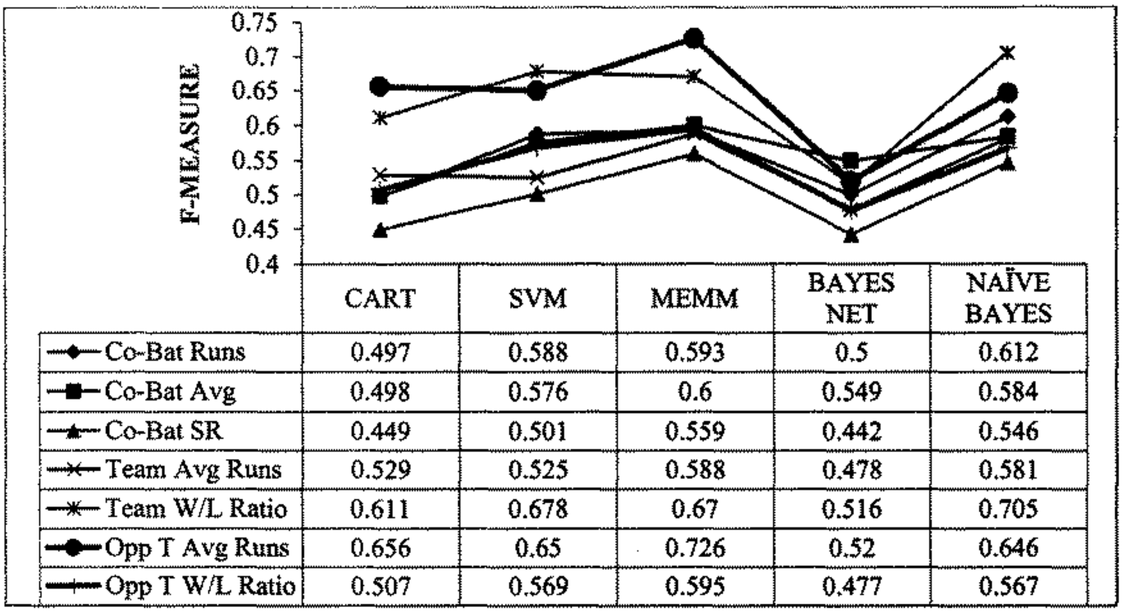
Figure 8 : Features F-Measure Analysis Using 1st Dataset (Bowlers)

Scenario 4: Individual Features Analysis Using Classification Models Based on ARI (Average Relative Increase) Dataset (Bowlers)

In this section we analyze f-measure of bowlers feature that are described in chapter 3 by using 2nd dataset of Average Relative Increase (ARI) of Runs in bowlers dataset.

Result: F-Measure Analysis of Features

Figure 9 show the f-measure result of features by using 2nd dataset (ARI) that consist of 10 to 100 players (bowlers). In figure 9, we note that Opposite Team Average Runs feature gives 65%, 65%, 72%, 52% and 64% accuracy result using CART, SVM, MEMM, BAYES NET and NAÏVE BAYES classifiers and Team Win Loss Ratio features gives second highest accuracy results.



	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Bat Runs	0.497	0.588	0.593	0.5	0.612
■ Co-Bat Avg	0.498	0.576	0.6	0.549	0.584
▲ Co-Bat SR	0.449	0.501	0.559	0.442	0.546
× Team Avg Runs	0.529	0.525	0.588	0.478	0.581
✳ Team W/L Ratio	0.611	0.678	0.67	0.516	0.705
● Opp T Avg Runs	0.656	0.65	0.726	0.52	0.646
+ Opp T W/L Ratio	0.507	0.569	0.595	0.477	0.567

Figure 9: Features F-Measure Analysis Using 2nd Dataset (Bowlers)

4.4.2 Model Wise Combined Features Analysis

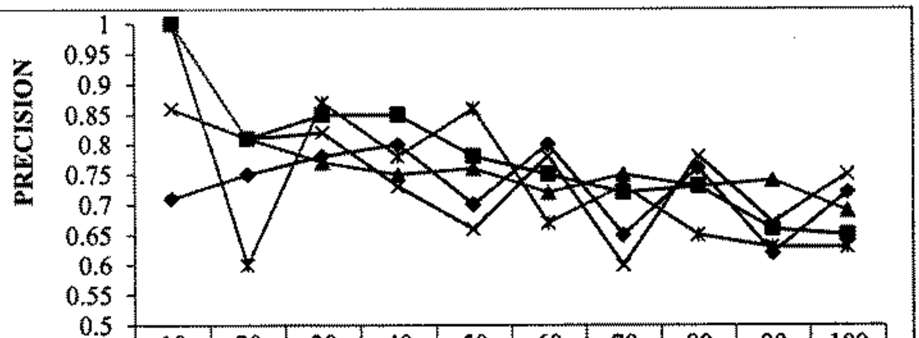
We have calculated precision, recall and f-measure of all the features by using classification models on two types of data for batsmen and bowlers. For both two type of data sample are selected, in the first type of data sample, players (batsmen and bowlers) are selected on the basis of players weighted average (of runs, average, and strike rate) and in the second type of data sample, players are selected for performance analysis on the basis of ARI (Average Relative Increase) notion. In model wise combined feature analysis we have selected all features at once to calculate precision, recall and f-measures and compared accuracy between different classification models that we have used.

Scenario 1: Combined Features Analysis Using Classification Models Based on Weighted Average Dataset (Batsmen)

In this section we analyzed precision, recall and f-measures of batsmen features that were described in chapter 3 by using classification models on two types if datasets. First row of table in figures show the data sample that consisted of 10 to 100 players and left column shows the models wise result.

Result: Model Wise Precision Analysis of Features

Figure 10 show the precision results of features by using five different classification models. The comparison of results between different classification models shows the performance of SVM model is better than other models that we have used. SVM Model gives the highest average accuracy result of 78%.

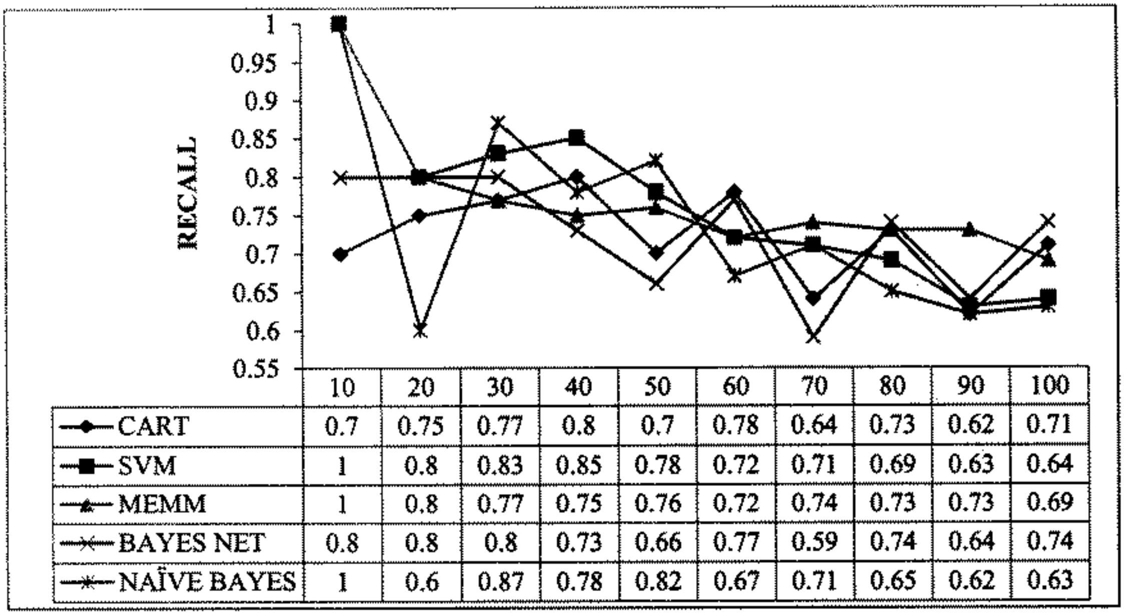


	10	20	30	40	50	60	70	80	90	100
—◆— CART	0.71	0.75	0.78	0.8	0.7	0.8	0.65	0.76	0.62	0.72
—■— SVM	1	0.81	0.85	0.85	0.78	0.75	0.72	0.73	0.66	0.65
—▲— MEMM	1	0.81	0.77	0.75	0.76	0.72	0.75	0.73	0.74	0.69
—✕— BAYES NET	0.86	0.81	0.82	0.73	0.66	0.78	0.6	0.78	0.67	0.75
—✱— NAÏVE BAYES	1	0.6	0.87	0.78	0.86	0.67	0.73	0.65	0.63	0.63

Figure 10: Model Wise Precision Analysis Using 1st Dataset (Batsmen)

Result: Model Wise Recall Analysis of Features

Similarly figure 11 show the recall result of features using classification models provide mostly the same result it gives in its precision in figure 10. Figure 11 shows that SVM and MEMM Models gives better result 76 % each.

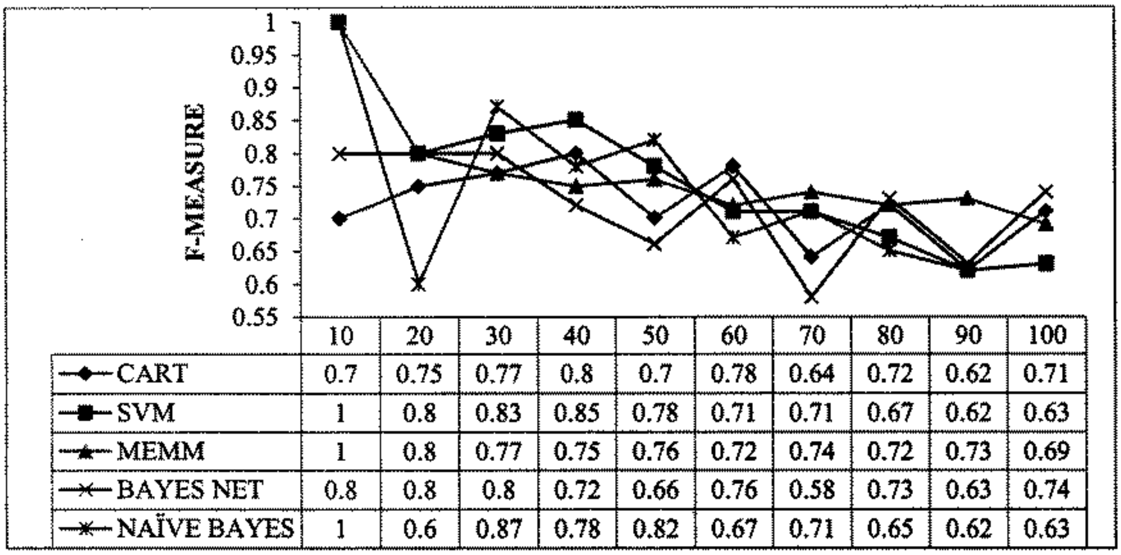


	10	20	30	40	50	60	70	80	90	100
◆ CART	0.7	0.75	0.77	0.8	0.7	0.78	0.64	0.73	0.62	0.71
■ SVM	1	0.8	0.83	0.85	0.78	0.72	0.71	0.69	0.63	0.64
▲ MEMM	1	0.8	0.77	0.75	0.76	0.72	0.74	0.73	0.73	0.69
× BAYES NET	0.8	0.8	0.8	0.73	0.66	0.77	0.59	0.74	0.64	0.74
✳ NAÏVE BAYES	1	0.6	0.87	0.78	0.82	0.67	0.71	0.65	0.62	0.63

Figure 11: Model Wise Recall Analysis Using 1st Dataset (Batsmen)

Result: Model Wise F-Measure Analysis of Features

Figure 12 show the f-measure results of features by using five different classification models. The comparison of results between different classification models shows the performance of SVM and MEMM model is better than other models that we have used. SVM and MEMM Models give the highest average accuracy result of 76%.



	10	20	30	40	50	60	70	80	90	100
◆ CART	0.7	0.75	0.77	0.8	0.7	0.78	0.64	0.72	0.62	0.71
■ SVM	1	0.8	0.83	0.85	0.78	0.71	0.71	0.67	0.62	0.63
▲ MEMM	1	0.8	0.77	0.75	0.76	0.72	0.74	0.72	0.73	0.69
× BAYES NET	0.8	0.8	0.8	0.72	0.66	0.76	0.58	0.73	0.63	0.74
✳ NAÏVE BAYES	1	0.6	0.87	0.78	0.82	0.67	0.71	0.65	0.62	0.63

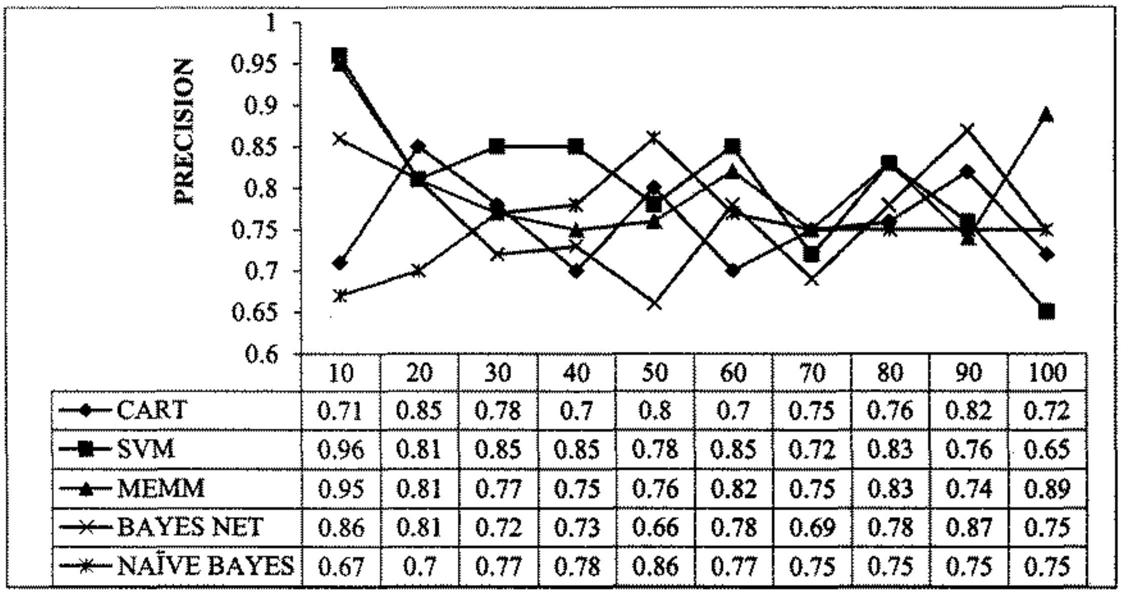
Figure 12: Model Wise F-Measure Analysis Using 1st Dataset (Batsmen)

Scenario 2: Combined Features Analysis Using Classification Models Based on ARI (Average Relative Increase) Dataset (Batsmen)

In this section we analyze f-measure of players (batsmen) based on 2nd dataset that is (ARI) Average Relative Increase.

Result: Model Wise Precision Analysis of Features

Figure 13 shows the precision results of features by using five different classification models on data sample that consist of 10 to 100 players (batsmen). In figure 13, we note that the performance of SVM and MEMM model is provide the 80 % which is the highest average accuracy result comparing with the other three models that are in the graph. That CART, BAYES NET and NAÏVE BAYES gives 75%, 76% and 75% average accuracy result respectively.

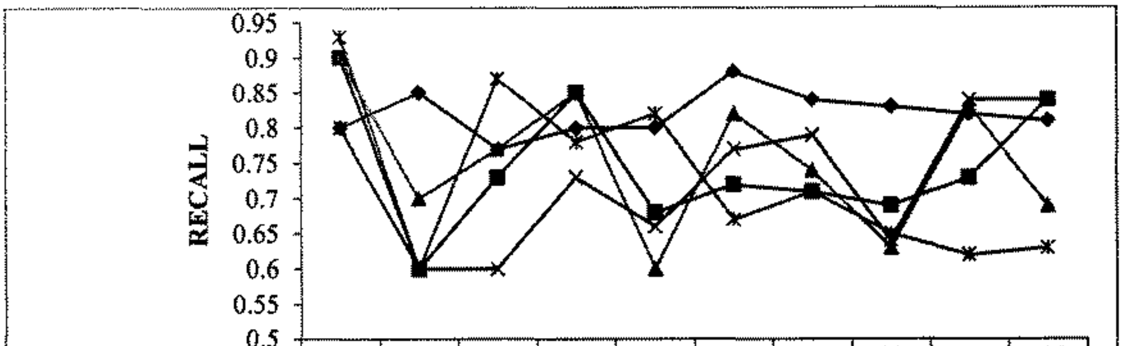


	10	20	30	40	50	60	70	80	90	100
◆ CART	0.71	0.85	0.78	0.7	0.8	0.7	0.75	0.76	0.82	0.72
■ SVM	0.96	0.81	0.85	0.85	0.78	0.85	0.72	0.83	0.76	0.65
▲ MEMM	0.95	0.81	0.77	0.75	0.76	0.82	0.75	0.83	0.74	0.89
✕ BAYES NET	0.86	0.81	0.72	0.73	0.66	0.78	0.69	0.78	0.87	0.75
✱ NAÏVE BAYES	0.67	0.7	0.77	0.78	0.86	0.77	0.75	0.75	0.75	0.75

Figure 13: Model Wise Precision Analysis Using 2nd Dataset (Batsmen)

Result: Model Wise Recall Analysis of Features

Figure 14 show the recall result of features of 2nd dataset (ARI) Average Relative Increase that consist of 10 to 100 players (batsmen) by using five classification models. In figure 14, we noted that models SVM, MEMM, BAYES NET and NAÏVE BAYES gives average result of accuracy are 82%, 74%, 75%, 72% and 72% respectively. Results show that CART model which gives 82 % result gives the highest average of accuracy among the all model that we used.

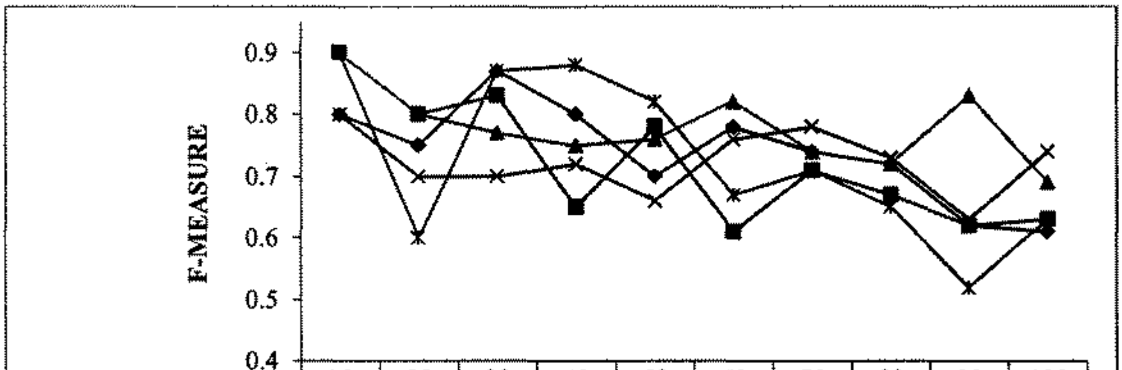


	10	20	30	40	50	60	70	80	90	100
◆ CART	0.8	0.85	0.77	0.8	0.8	0.88	0.84	0.83	0.82	0.81
■ SVM	0.9	0.6	0.73	0.85	0.68	0.72	0.71	0.69	0.73	0.84
▲ MEMM	0.9	0.7	0.77	0.85	0.6	0.82	0.74	0.63	0.83	0.69
× BAYES NET	0.8	0.6	0.6	0.73	0.66	0.77	0.79	0.64	0.84	0.84
✳ NAÏVE BAYES	0.93	0.6	0.87	0.78	0.82	0.67	0.71	0.65	0.62	0.63

Figure 14: Model Wise Recall Analysis Using 2nd Dataset (Batsmen)

Result: Model Wise F-Measure Analysis of Features

Similarly figure 15 shows the f-measure result of features using classification models. In figure 15, we note that overall all models give better result because result shows that each model give result in above 70 percent but MEMM model provide the highest accuracy model which is 77%.



	10	20	30	40	50	60	70	80	90	100
◆ CART	0.8	0.75	0.87	0.8	0.7	0.78	0.74	0.72	0.62	0.61
■ SVM	0.9	0.8	0.83	0.65	0.78	0.61	0.71	0.67	0.62	0.63
▲ MEMM	0.9	0.8	0.77	0.75	0.76	0.82	0.74	0.72	0.83	0.69
× BAYES NET	0.8	0.7	0.7	0.72	0.66	0.76	0.78	0.73	0.63	0.74
✳ NAÏVE BAYES	0.9	0.6	0.87	0.88	0.82	0.67	0.71	0.65	0.52	0.63

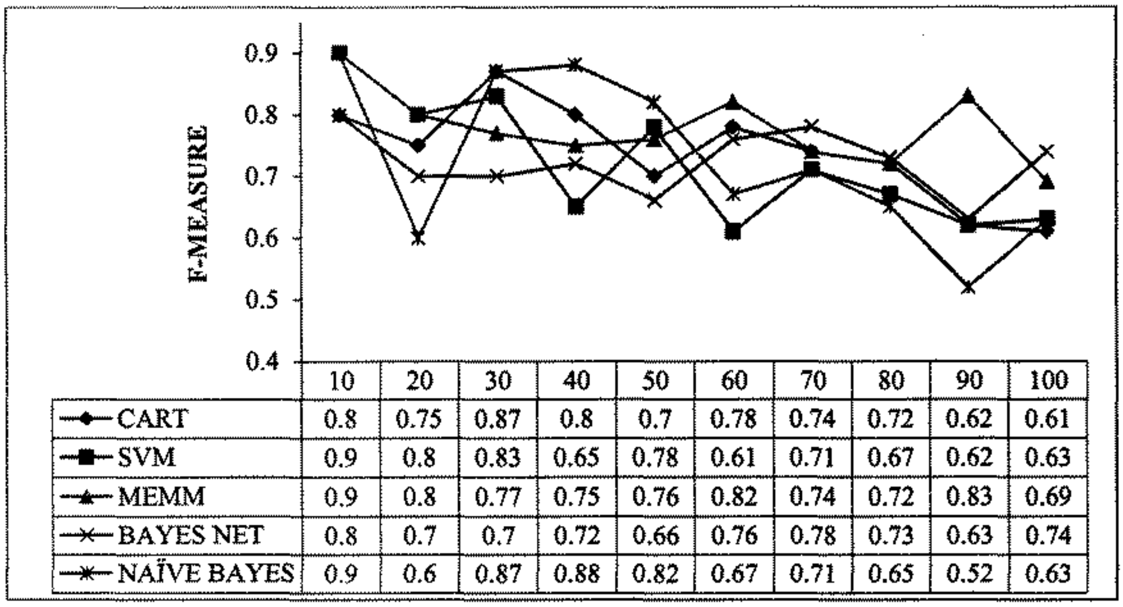
Figure 15: Model Wise F-Measure Analysis Using 2nd Dataset (Batsmen)

Scenario 3: Combined Features Analysis Using Classification Models Based on Weighted Average Dataset (Bowlers)

In this section we analyzed f-measure of players (bowlers) features that were describe in detail in chapter 3 by using classification models. First rows show the data samples that consisted of 10 to 100 player and all other row shows the f-measure results.

Result: Model Wise F-Measure Analysis of Features

Figure 16 show the f-measure results of features by using five different classification models. The comparison of results between different classification models shows the performance of SVM model is better than other models that we have used. SVM Model gives the highest average accuracy result of 81%. While the other models gives result are 79%, 77%, 79% and 76% in a sequence shows in figure 16.



	10	20	30	40	50	60	70	80	90	100
—◆— CART	0.8	0.75	0.87	0.8	0.7	0.78	0.74	0.72	0.62	0.61
—■— SVM	0.9	0.8	0.83	0.65	0.78	0.61	0.71	0.67	0.62	0.63
—▲— MEMM	0.9	0.8	0.77	0.75	0.76	0.82	0.74	0.72	0.83	0.69
—×— BAYES NET	0.8	0.7	0.7	0.72	0.66	0.76	0.78	0.73	0.63	0.74
—✱— NAÏVE BAYES	0.9	0.6	0.87	0.88	0.82	0.67	0.71	0.65	0.52	0.63

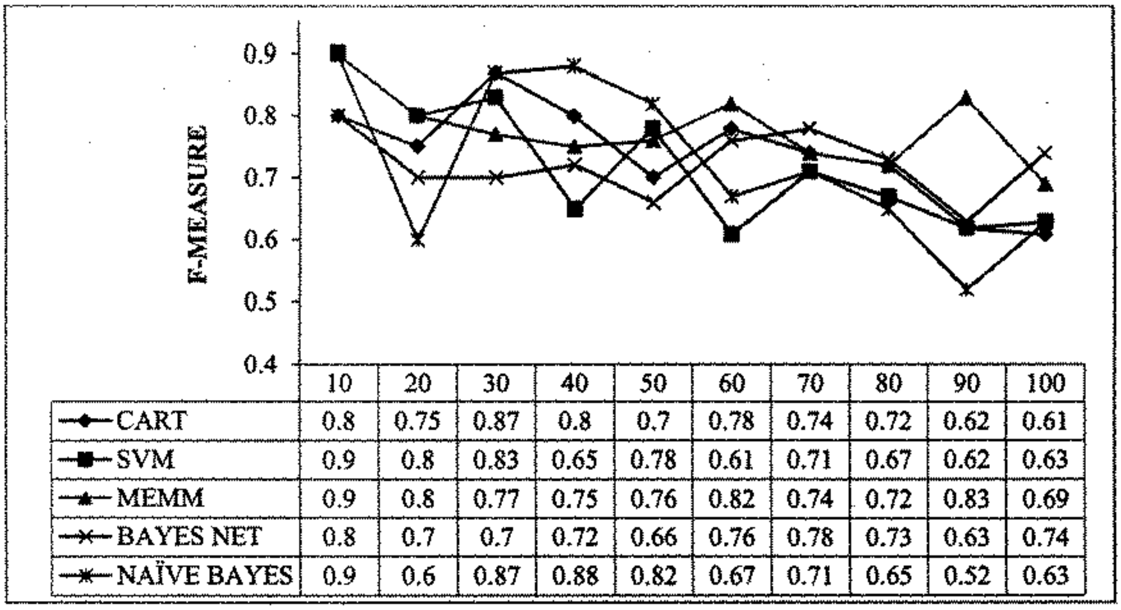
Figure 16: Model Wise F-Measure Analysis Using 1st Dataset (Bowlers)

Scenario 4: Combined Features Analysis Using Classification Models Based on ARI (Average Relative Increase) Dataset (Bowlers)

We analyzed f-measure of players (bowlers) features in this section that were describe in detail in chapter 3 by using classification models on 2nd type of data set that is (ARI) Average relative Increase in wicket here because the dataset of player bowlers are used.

Result: Model Wise F-Measure Analysis of Features

Figure 17 show the f-measure result of features of 2nd dataset (ARI) Average Relative Increase that consist of 10 to 100 players (batsmen) by using five classification models. In figure 17, we noted that models SVM, MEMM, BAYES NET and NAÏVE BAYES gives average result of accuracy are 83%, 79%, 79%, 22% and 80% respectively. Results show that CART model which gives 83 % result gives the highest average of accuracy among the all model that we used.



	10	20	30	40	50	60	70	80	90	100
—◆—CART	0.8	0.75	0.87	0.8	0.7	0.78	0.74	0.72	0.62	0.61
—■—SVM	0.9	0.8	0.83	0.65	0.78	0.61	0.71	0.67	0.62	0.63
—▲—MEMM	0.9	0.8	0.77	0.75	0.76	0.82	0.74	0.72	0.83	0.69
—×—BAYES NET	0.8	0.7	0.7	0.72	0.66	0.76	0.78	0.73	0.63	0.74
—✳—NAÏVE BAYES	0.9	0.6	0.87	0.88	0.82	0.67	0.71	0.65	0.52	0.63

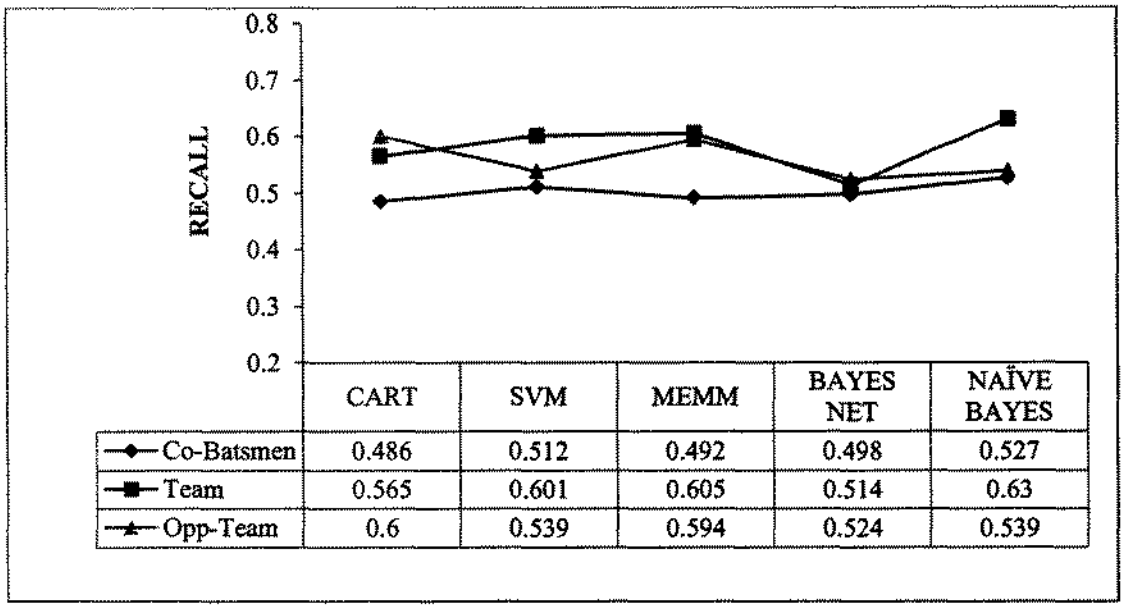
Figure 17: Model Wise F-Measure Analysis Using 2nd Dataset (Bowlers)

4.4.3 Category Wise Analysis

For this section, first we have categorized our features in three groups such as co-batsmen (Runs, Average, Strike Rate), Team (Team Win Loss Ratio and Team Average Runs), Opposite Team (Opposite Team Win Loss Ratio and Opposite Team Average Runs). Then we calculate the precision, recall and f-measures of features groups by using CART, SVM, MEMM, BAYES NET and NAÏVE BAYES models. Two types of dataset have used; first type of dataset is based

Result: Recall Analysis of Features Categories

Similarly figure 19 show the recall result of features categories using classification models provide mostly the same result it gives in its precision in figure 18. Figure 19 shows that team category gives better result using NAVIE BAYES model give 63%, while opposite team category give highest average accuracy result using CART and co-batsmen gives its highest result on model NAÏVE BAYES they give 60 % and 52% respectively

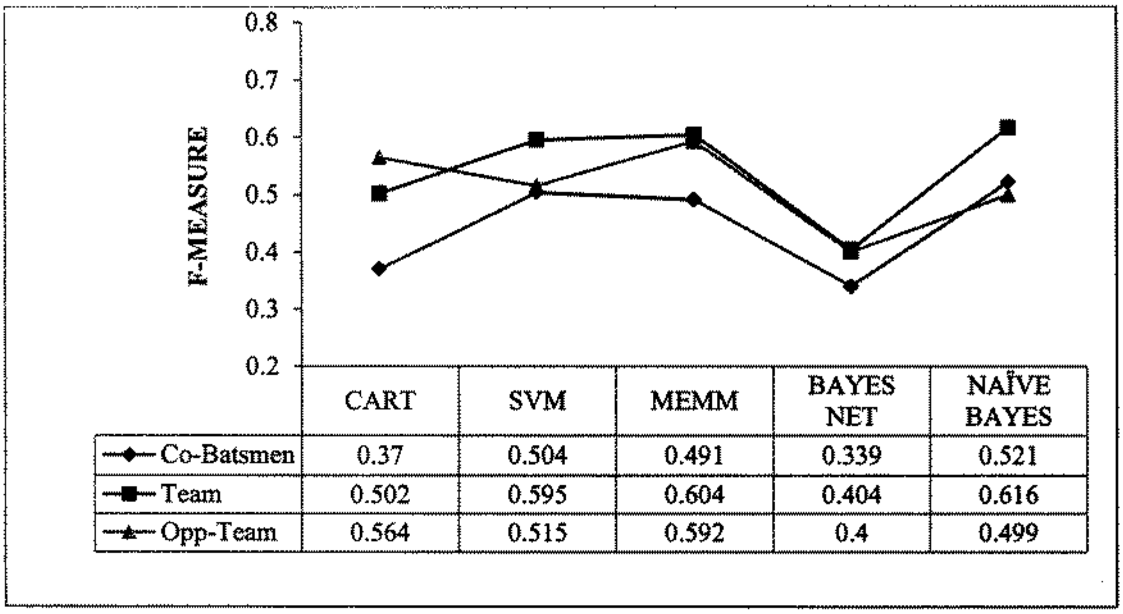


	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Batsmen	0.486	0.512	0.492	0.498	0.527
■ Team	0.565	0.601	0.605	0.514	0.63
▲ Opp-Team	0.6	0.539	0.594	0.524	0.539

Figure 19: Category Wise Recall Analysis Using 1st Dataset (Batsmen)

Result: F-Measure Analysis of Features Categories

Figure 20 shows the f-measure result of features categories using classification models. In figure 20, we note that f-measure result is almost same to its recall analysis show in figure 19. Here team category provides highest average accuracy result which is 61% using NAÏVE BAYES model.



	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Batsmen	0.37	0.504	0.491	0.339	0.521
■ Team	0.502	0.595	0.604	0.404	0.616
▲ Opp-Team	0.564	0.515	0.592	0.4	0.499

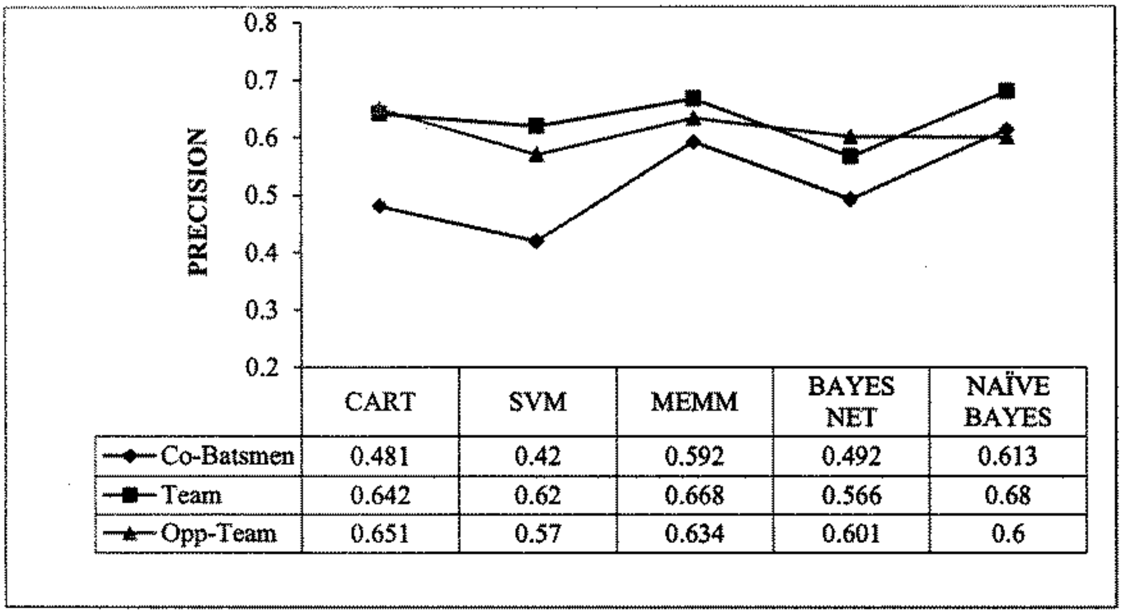
Figure 20: Category Wise F-Measure Analysis Using 1st Dataset (Batsmen)

Scenario 2: Features Categories Analysis Using Classification Models Based on ARI (Average Relative Increase) Dataset (Batsmen)

We analyzed precision, recall and f-measure of players (batsmen) features categories in this section that were describe in detail in chapter 3 by using classification models on 2nd type of data set that is (ARI) Average relative Increase of runs here because the dataset of player batsmen are used

Result: Precision Analysis of Features Categories

Figure 21 show the precision results of features categories by using classification models on the data sample that consist of 10 to 100 players. The comparison results among the features category shows that the performance of Team category gives 66% highest accuracy using MEMM model is better than Co-Batsmen that provides 61% highest accuracy using NAÏVE BAYES and Opposite Team that provide 65% highest accuracy using CART model.

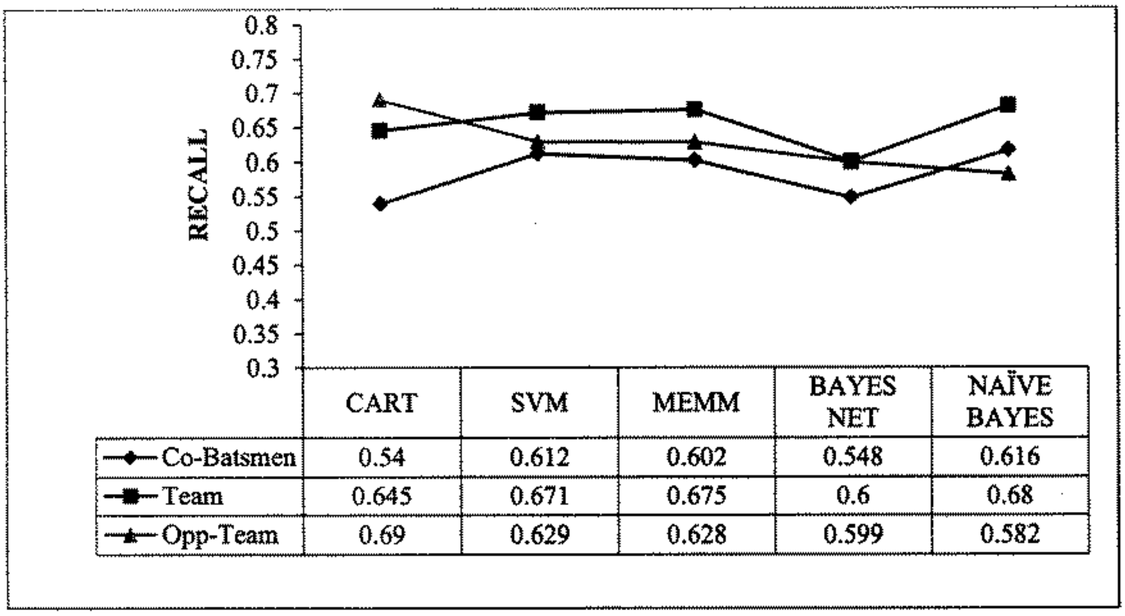


	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Batsmen	0.481	0.42	0.592	0.492	0.613
■ Team	0.642	0.62	0.668	0.566	0.68
▲ Opp-Team	0.651	0.57	0.634	0.601	0.6

Figure 21: Category Wise Precision Analysis Using 2nd Dataset (Batsmen)

Result: Recall Analysis of Features Categories

Similarly figure 22 shows the recall result of features using classification models. In figure 15, we note that overall all models give better result because result shows that each model gives approximately 60% percent or above but category team using MEMM model provide the highest accuracy model which is 67%.

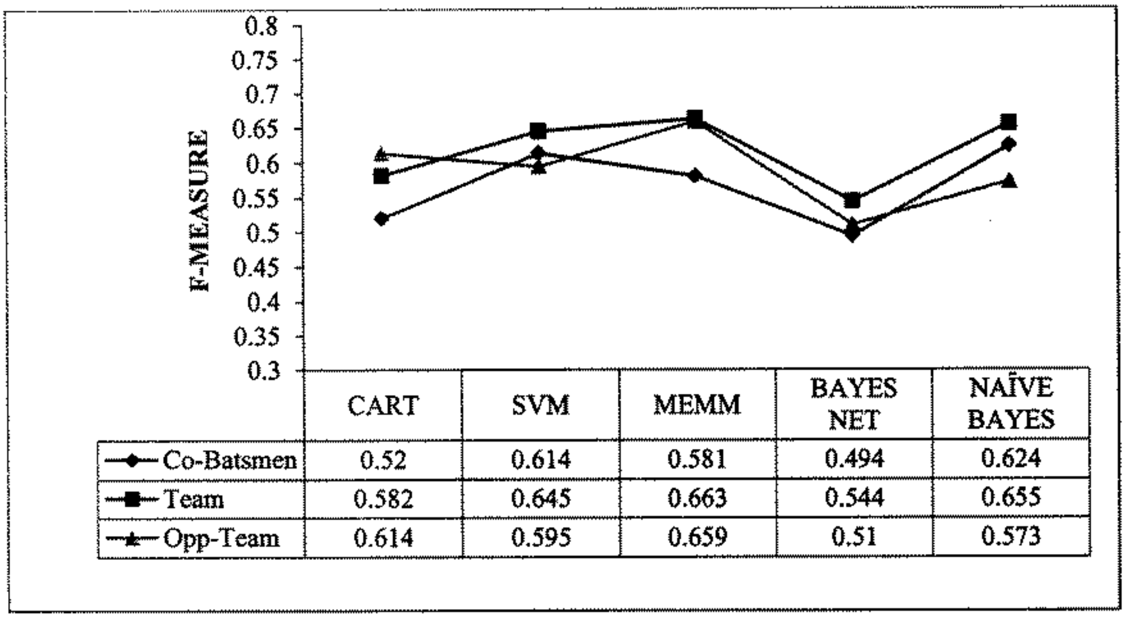


	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Batsmen	0.54	0.612	0.602	0.548	0.616
■ Team	0.645	0.671	0.675	0.6	0.68
▲ Opp-Team	0.69	0.629	0.628	0.599	0.582

Figure 22: Category Wise Recall Analysis Using 2nd Dataset (Batsmen)

Result: F-Measure Analysis of Features Categories

Similarly figure 23 show the f-measure result of features categories on 2nd type of dataset using classification models. Figure 23 shows that team category gives better result using MEMM model give 66%, while opposite team category give highest average accuracy result using same model (MEMM) and co-batsmen gives its highest result on model NAÏVE BAYES they give 65% and 62% respectively



	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
—◆— Co-Batsmen	0.52	0.614	0.581	0.494	0.624
—■— Team	0.582	0.645	0.663	0.544	0.655
—▲— Opp-Team	0.614	0.595	0.659	0.51	0.573

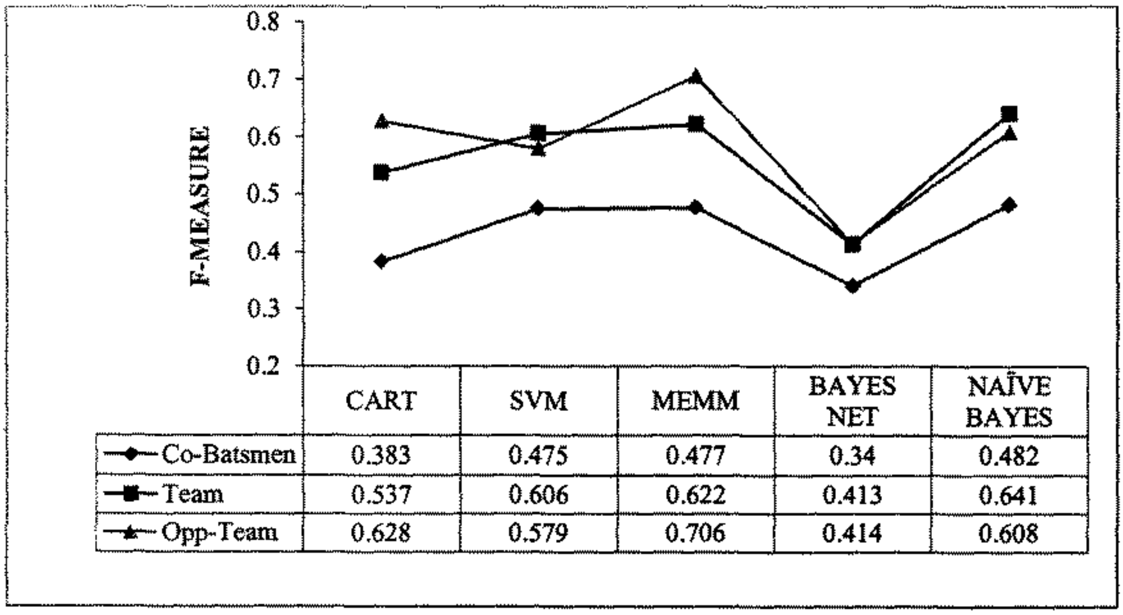
Figure 23: Category Wise F-Measure Analysis Using 2nd Dataset (Batsmen)

Scenario 3: Features Categories Analysis Using Classification Models Based on Weighted Average Dataset (Bowlers)

In this section we analyze f-measure of bowler's features categories by using 1st dataset which is based on Weighted Average (runs, average and strike rate) described in the last section of chapter 3 on data sample of 10 to 100 players (bowlers).

Result: F-Measure Analysis of Features Categories

Figure 24 shows the f-measure result of features categories using classification models. In figure 24, we note that the comparison results among the features category shows that the performance of Opposite Team category gives 70% highest accuracy using MEMM model is better than Co-Batsmen that provides 47% highest accuracy using MEMM and Team that provide 64% highest accuracy using NAÏVE BAYES model.



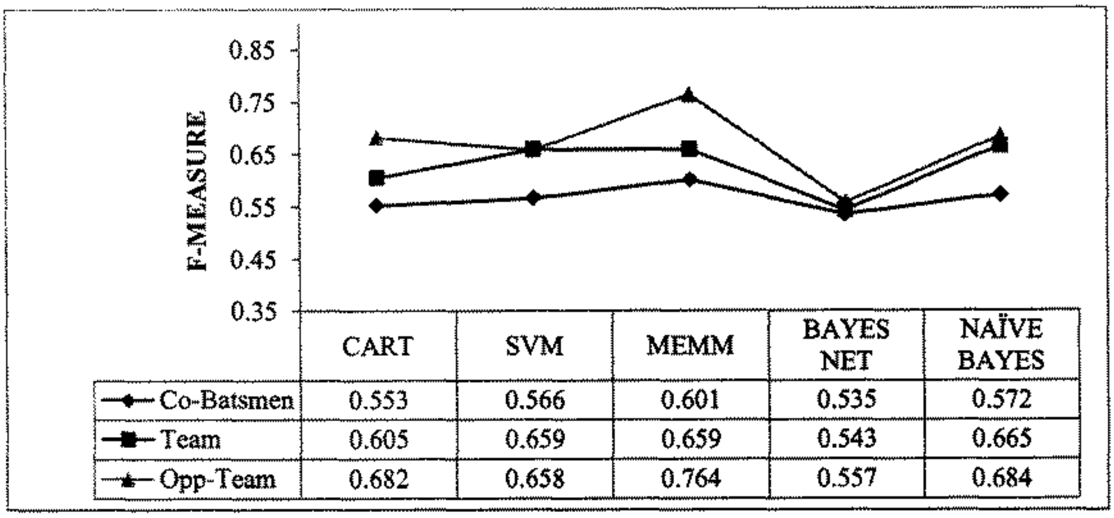
	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
◆ Co-Batsmen	0.383	0.475	0.477	0.34	0.482
■ Team	0.537	0.606	0.622	0.413	0.641
▲ Opp-Team	0.628	0.579	0.706	0.414	0.608

Figure 24: Category Wise F-Measure Analysis Using 1st Dataset (Bowlers)

Scenario 4: Features Categories Analysis Using Classification Models Based on ARI (Average Relative Increase) Dataset (Bowlers)

Result: F-Measure Analysis of Features Categories

The f-measure result of features by using classification models on 2nd dataset that is (ARI) Average Relative increase of wickets. Graph in figure 24 shows that category Opposite Team shows highest position with the highest average accuracy of 76% by using MEMM model. Team category provides the second highest average accuracy result by using SVM and MEMM which is 65%.



	CART	SVM	MEMM	BAYES NET	NAÏVE BAYES
Co-Batsmen	0.553	0.566	0.601	0.535	0.572
Team	0.605	0.659	0.659	0.543	0.665
Opp-Team	0.682	0.658	0.764	0.557	0.684

Figure 25: Category Wise F-Measure Analysis Using 2nd Dataset (Bowlers)

4.5 Rising Stars

We have predict rising stars using two different techniques that are described below in detail with rising stars prediction results. We have used these techniques on both dataset of batsmen and bowlers.

4.5.1 Rising Star Score (Batsmen)

Top ten rising stars found using rising stars score that are shown in table 9 and table 10. We have searched their ranking (www.cricinfo.org) in 2010 to compare our predicted result with player ranking in 2010. Table shows the Player name with player ranking data in year 2011. First we take the sum of co-batsmen average, co-batsmen strike rate, co-batsmen runs, team win loss ratio, team average, opposite team win loss ratio and opposite team average as show in the following expression.

Rising Star Score (Batsmen) = Co Bat Avg + Co Bat SR + Co Bat Runs + Team Win Loss Ratio + Team Avg + Opposite Team Win Loss Ratio + Opp Team Avg

Finally, we sort in descending order rising star score of each player then we took top ten highest rising star score player (Batsmen).

Table 9 : Top Ten Predicted Rising Star (Batsmen) From Rising Star Score

S No	Players	Country	Rising Star Score	Ranking in 2011
1	V Kohli	India	2609.56	5
2	SE Marsh	Australia	2575.83	35
3	RG Sharma	India	2490.06	42
4	Umar Akmal	Pakistan	2335.74	24
5	HM Amla	South Africa	2247.75	1
6	CL White	Australia	2247.11	38
7	SK Raina	India	2103.46	40
8	SR Watson	Australia	2039.25	12
9	Salman Butt	Pakistan	2027.87	36
10	MJ Clake	Australia	2018.41	6

Table 9 shows the top ranked player names, country for which he played, rising star score and their ranking in 2011 as per average take from www.cricinfo.org/statsguru. V Kohli ranking 1st in our predicted rising star score and as per ranking in 2011 he had ranked 5. Serial number which is shown in table 9 most left column is the rising star score ranking, while right most column in table 9 show player ranking in year 2011 which make easy to analyze that how much is the predicted rising star are correct.

4.5.2 Average Relative Increase (Batsmen)

Second predicted ranking list of top ten players (batsmen) for predicted rising stars using 2nd dataset is presented in table 10. The mathematically formulation of ARI score is describe in section 4.1.

Table 10: Top Ten Predicted Rising Star (Batsmen) From ARI Method

S No	Players	Country	ARI Score	Ranking in 2011
1	KD Karthik	India	15785.28	68
2	RG Sharma	India	1021.311	42
3	CK Coventry	Zimbabwe	951.1111	>100
4	AB de Villiers	South Africa	505.4417	8
5	JM How	New Zealand	435.6316	>100
6	SK Raina	India	390.5523	40
7	SC Williams	Zimbabwe	372.4954	>100
8	RR Sarwan	West Indies	237.593	19
9	M. Ashraful	Bangladesh	236.039	79
10	MJ Clake	Australia	198.5462	6

Player name, country, ARI (Average Relative Increase) Score and ranking in 2011 shows in table 10. Table shows the serial 3, 5, and 7 the players CK Coventry, JM How and SC Williams respectively. After the calculation we got the top 10 predicted players using ARI (Average Relative Increase) score. Up on searching for ranking I couldn't find any ranking of these player or they may be after 100th position, so we simple put >100 in place of these players ranking, which mean greater than hundred (>100).

4.5.3 Rising Star Score (Bowler)

In this section we predict rising bowler using two main methods which are discussed in chapter 3 and chapter 4. First method, which is predicted rising stars with rising star score for in which we sum all the features that we discussed in chapter 3 and got the rising star score. In Second method, this is ARI of wickets for bowlers. Following expression shows the Rising Star Score for bowlers.

Rising Star Score (Bowlers) = Co Bowl Avg + Co Bowl SR + Co Bowl Econ + Team Win Loss Ratio + Team Avg + Opposite Team Win Loss Ratio + Opp Team Avg

After adding all the feature we got rising star score for each individual player than we sort in descending order rising star score of each player then we took top ten highest rising star score player (Bowler). Following table shows the top ten predicted rising star using rising star score.

Table 11 : Top Ten Predicted Rising Star (Bowlers) From Rising Star Score

S No	Players	Country	Rising Star Score	ICC Ranking in 2011
1	Mashrafe Mortaza	Bangladesh	4.36	15
2	Shakib Al Hasan	Bangladesh	3.95	6
3	Z Khan	India	3.94	13
4	A Nehra	India	3.90	25
5	B Lee	India	3.90	1
6	Harbajan Singh	India	3.87	12
7	NW Braken	Australia	3.84	9
8	RP Singh	India	3.84	46
9	I Sharma	India	3.83	53
10	M Morkel	South Africa	3.81	37

Table 11 shows the top ranked player names, country for which he played, rising star score and their ranking in 2011 as per average take from www.cricinfo.org/statsguru. Mashrafe Mortaza from Bangladesh ranking 1st in our predicted rising star score and as per ranking in 2011 he had ranked 15. Similarly Z Khan (Zaheer Khan) from India is placed 3 in our predicted top ten rising star bowling and he ranked 13 in 2011. One of the best predictions in the table 11 is predicted player who ranked 5 in our predicted table while he was ranked 1 in the 2011.

4.5.4 Average Relative Increase (Bowlers)

Following table shows top ten predicted rising star using ARI (Average Relative Increase) score. After getting ARI of each player first sort it in descending with highest ARI player from top to bottom than we selected top ten players as out top ten predicted rising star bowler following table shows that top ten rising stars.

Table 12 : Top Ten Predicted Rising Star (Bowlers) From ARI Score

S No	Players	Country	ARI Score	Ranking in 2011
1	Naveed-ul-Hassan	Pakistan	1019.89	27
2	SL Malinga	Sri Lanka	1018.25	7
3	CK Langeeveldt	South Africa	945.48	29
4	Mahmudullah	Bangladesh	800.00	90
5	DW Steyn	South Africa	571.43	29
6	TT Bresnan	England	550.00	47
7	J Botha	South Africa	344.44	42
8	GP Swan	England	328.57	28
9	JS Patel	New Zealand	314.29	93
10	KMDN Kulasekara	Sri Lanka	302.78	20

Player name, country, ARI (Average Relative Increase) Score and ranking in 2011 shows in table 12. Table 12 shows the comparison between serial columns which is our predicted ranked players with player ranking in 2011 column. Like for instance SL Malinga's rank in our predicted table is 2 and in ranking from www.cricinfo.org/statsguru, he ranked 7 in 2011.

CHAPTER 5

CONCLUSION

5. CONCLUSION

Predicting rising stars by using classification modeling techniques is a difficult human resource exercise. As discussed above that no such work is so far done in cricket for predicting players using classification modeling techniques. But it has been done recently in different domain like in research communities, which generate and predict real good numbers. So we use that technique of predicting rising stars in research community using classification model in the cricket domain.

First we collect player's datasets from a web (www.espncricinfo.org/stasts/statsguru) of year 2000 to 2009 then we have constructed co-player datasets separately for batsmen and bowlers. After that we describe distinct features categories that are co-player, team and opposite team which are discussed in detail in chapter 3. We have used five classification models (CART, SVM, MEMM, BAYES NET and NAÏVE BAYES) on that dataset compare the result with each other. Two type of classification models are used, such as discriminative and generative on our two types of dataset for bowling and batting. In short about classification model results, discriminative (CART, SVM and MEMM) models provide higher accuracy as compare to generative (BAYES NET and NAÏVE BAYES) models.

Finally, we use two methods for predicting rising stars which are rising star score and Average Relative Increase (ARI). Both methods are discussed in section 4.5.1 and 4.5.2. Top ten our predicted players (batsmen/bowlers) in 2009 shown in table in last section of chapter 5 having players ranking of year 2011 for comparing our prediction. Some of predicted player ranked in top ten later on.

REFERENCES

ESPNcricinfo. (n.d.). Retrieved from ESPNcricinfo: http://www.espncricinfo.com/

CricketArchive. (n.d.). Retrieved from CricketArchive: http://cricketarchive.com/

Wikipedia. http://en.wikipedia.org/wiki/List_of_International_Cricket_Council_members

S.Mukherjee, "Quantifying individual performance in Cricket - A network analysis of Batsmen and Bowlers", Physica A, 393, 624, 2013.

P.Kalgotra, R.Sharda and G.Chakraborty, "Predictive Modeling in Sports Leagues: An Application in Indian Premier League", in Proceedings of the SAS Global forum, 2013.

S.Mukherjee," Identifying the greatest team and captain - A complex network approach to cricket matches", Physica A 391, 23, 6066 , 2012.

P.Lucey, A.Bialkowski, M. Monfort, P.Carr and L.Matthews ""Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data", Disney Research Pittsburgh, PA, USA, 15232, 2015.

A.C. Constantinou, N.E. Fenton, M. Neil, "pi-football: a Bayesian network model for forecasting association football match outcomes", Knowledge-Based Systems 36 322–339, 2012.

Mascaro, S., Nicholso, A., & Korb, K, "Anomaly detection in vessel tracks using Bayesian networks", *International Journal of Approximate Reasoning, Vol. 1, No. 1*,(pp. 84-98), 2014.

Daud, A., Abbasi, R., & Muhammad, F, "Finding Rising Stars in Social Networks" *Springer, Vol 7825*, (pp. 13-24), 2013.

Bolivar, A. R., Hidrobo, F., & G, P, "Fault Diagnosis In Dynamic Processes: A Data Mining and SVM Application", *International Journal of Systems Applications, Engineering & Development*, (pp. 191-199), 2013

Page, K. L., Ward, K., & Worrall-Carter, "The Process and Utility of Classification and Regression Tree Methodology in nursing resaerch", *Journal of Advance Nursing*, 2013.

Mitchell, T. M, "Generative and Discriminative Classifiers: Naive Bayes And Logistic Regression", In T. M. Mitchell, & M. Hill, *Machine Learning* (pp. 1-17), 2010.

Marcos, Z, "Using Bag-of-words to Distinguish Similar Languages: How Efficient are They? *Computational Intelligence and Informatics (CINTI)", IEEE 14th International Symposium,* (pp. 37-41), 2013.

R.Stuckless, "Developments in real-time speech-to-text communication for people with impaired hearing", Communication Access for People with Hearing Loss, York Press, Baltimore, MD, pp 197226, 1994.

H.Saikia & D.Bhattacharjee, "On Classification of All-rounders of the Indian Premier League (IPL): A Bayesian Approach", Vikalpa, Vol.36 No.4 (pp. 25 - 40), 2011.

Y. Yue, P. Lucey, P. Carr, A. Bialkowski and I. Matthews, "Learning Fine-Grained Spatial Models for Dynamic Sports Play Prediction", in the International Conference of Data Mining (ICDM), 2014.

H.Saikia & D.Bhattacharjee, "A Bayesian Classification Model for Predicting the Performance of All-rounders in the Indian Premier League", http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1622060, 2010

Y.Fang, L.Si, & A.P.Mathur, "Discriminative models of integrating document evidence and document-candidate associations for expert search", SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 683 - 690). ACM, 2010.

T.Y.Liu, J.Xu, T.Qin, W.Xiong & H.Li, "Benchmark dataset for research on learning to rank for information retrieval", In SIGIR Workshop on Learning to Rank for Information Retrieval, (pp. 3 - 10), 2007.

J.Han and M.Kamber, "Data Mining Concepts and Techniques", TheMorgan Kaufmann Series in DataManagement Systems, 2006.

R.Nallapati," Discriminative models for information retrieval", in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 64 - 71). New York: ACM, 2004.

J.Tang, M.Hong, D.Zhang,B.Liang and J.Li, "Information Extraction: Methodologies and Applications", In H. A. Prado, & E. Ferneda, Emerging Technologies of Text Mining: Techniques and Applications (pp. 1 - 38). Beijing, China: IGI Global, 2007.

X.L.Li., C.S.Foo, K.L.Tew, and S.K.Ng, "Searching for Rising Stars in Bibliography Networks", DASFAA '09 Proceedings of the 14th International Conference on Database Systems for Advanced Applications (pp. 288 - 292). Springer-Verlag Berlin, Heidelberg, 2009.

Rising Cricketer Prediction Using Classification Models

Waqas Ahmad

665-FBAS/MSCS/S12

Department of Computer Science & Software Engineering

Faculty of Basic and Applied Sciences

International Islamic University Islamabad

2015